COHORT SQUEEZE: BEYOND A SINGLE COMMUNICATION ROUND PER COHORT IN CROSS-DEVICE FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtually all federated learning (FL) methods, including FedAvg, operate in the following manner: i) an orchestrating server sends the current model parameters to a cohort of clients selected via certain rule, ii) these clients then independently perform a local training procedure (e.g., via SGD or Adam) using their own training data, and iii) the resulting models are shipped to the server for aggregation. This process is repeated until a model of suitable quality is found. A notable feature of these methods is that each cohort is involved in a single communication round with the server only. In this work we challenge this algorithmic design primitive and investigate whether it is possible to “squeeze more juice” out of each cohort than what is possible in a single communication round. Surprisingly, we find that this is indeed the case, and our approach leads to up to 74% reduction in the total communication cost in standard cross-device FL, and up to 95% reduction in a hierarchical FL deployment. Our method is based on a novel variant of the stochastic proximal point method (SPPM-AS) which supports a large collection of client sampling procedures some of which lead to further gains when compared to classical client selection approaches.

1 INTRODUCTION

Federated Learning (FL) is increasingly recognized for its ability to enable collaborative training of a global model across heterogeneous clients, while preserving privacy (McMahan et al., 2016; 2017; Kairouz et al., 2019; Li et al., 2020a; Karimireddy et al., 2020b; Mishchenko et al., 2022b; Malinovsky et al., 2024; Yi et al., 2024). This approach is particularly noteworthy in cross-device FL, involving the coordination of millions of mobile devices by a central server for training purposes (Kairouz et al., 2019). This setting is characterized by intermittent connectivity and limited resources. Consequently, only a subset of client devices participates in each communication round. Typically, the server samples a batch of clients (referred to as a *cohort* in FL), and each selected client trains the model received from the server using its local data. Then, the server aggregates the results sent from the selected cohort. Another notable limitation of this approach is the constraint that prevents workers from storing states (operating in a stateless regime), thereby eliminating the possibility of employing variance reduction techniques.

We will consider a reformulation of the cross-device objective that assumes a finite number of workers being selected with uniform probabilities. Given that, in practice, only a finite number of devices is considered, i.e. the following finite-sum objective is considered:

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x). \tag{1}$$

This reformulation aligns more closely with empirical observations and enhances understanding for illustrative purposes. The extension to the expectation form of the following theory can be found in Appendix G.4.

Current representative approaches in the cross-device setting include FedAvg and FedProx. In our work, we introduce a method by generalizing stochastic proximal point method with arbitray sampling and term as SPPM-AS. This new method is inspired by the stochastic proximal point method

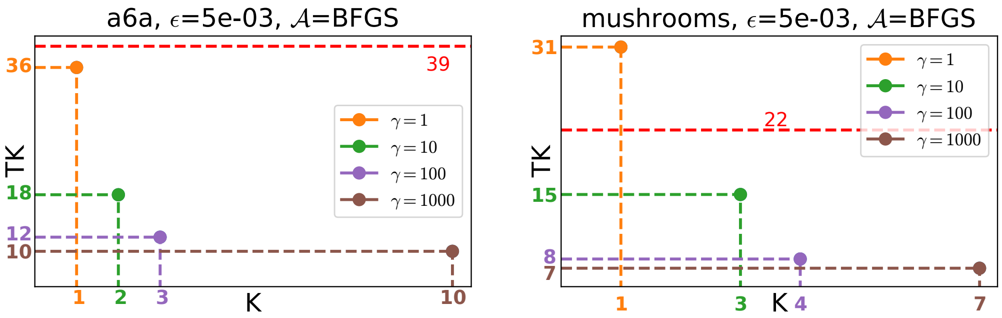


Figure 1: Total communication cost TK versus number of local communication rounds K required to reach ε-accuracy for logistic regression on `a6a` and `mushrooms` with cohort size $|C| = 10$. The dashed red line is FedAvg/LocalGD with $K = 1$.

(SPPM), a technique notable for its ability to converge under arbitrarily large learning rates and its flexibility in incorporating various solvers to perform proximal steps. This adaptability makes SPPM highly suitable for cross-device FL (Li et al., 2020a; Yuan & Li, 2022; 2023; Khaled & Jin, 2023; Lin et al., 2024). Additionally, we introduce support for an arbitrary cohort sampling strategy, accompanied by a theoretical analysis. We present novel strategies that include support for client clustering, which demonstrate both theoretical and practical improvements.

Another interesting parameter that allows for control is the number of local communications. Two distinct types of communication, *global* and *local*, are considered. A *global* iteration is defined as a single round of communication between the server and all participating clients. On the other hand, *local* communication rounds are synchronizations that take place within a chosen cohort. Additionally, we introduce the concept of total communication cost, which includes both local and global communication iterations, to measure the overall efficiency of the communication process. The total communication cost naturally depends on several factors. These include the local algorithm used to calculate the prox, the global stepsize, and the sampling technique.

Previous results on cross-device settings consider only one local communication round for the selected cohort (Li et al., 2020b; Reddi et al., 2020; Li et al., 2020a; Wang et al., 2021a;b; Xu et al., 2021; Malinovsky et al., 2023; Jhunjhunwala et al., 2023; Sun et al., 2023; 2024). Our experimental findings reveal that *increasing the number of local communication rounds within a chosen cohort per global iteration can indeed lower the total communication cost needed to reach a desired global accuracy level*, which we denote as ε. Figure 1 illustrates the relationship between total communication costs and the number of local communication rounds. In this figure, we use ℓ_2-regularized logistic regression on the LibSVM datasets `a6a` and `mushrooms` with cohort size $|C| = 10$. The proximal subproblem is solved by BFGS, and we report the total communication cost TK required to reach $\varepsilon = 5 \times 10^{-3}$ in $\|x_T - x^\star\|^2$, as detailed in Sec. 3.3 and App. E.1. Assume that the cost of communication per round is 1 unit. K represents the number of local communication rounds per global iteration for the selected cohort, while T signifies the *minimum* number of global iterations needed to achieve the accuracy threshold ϵ. Then, the total cost incurred by our method can be expressed as TK. For comparison, the dashed line in the figure shows the total cost for the FedAvg algorithm, which always sets K to 1, directly equating the number of global iterations to total costs. Our results across various datasets identify the optimal K for each learning rate to achieve ϵ-accuracy. Figure 1 shows that adding more local communication rounds within each global iteration can lead to a significant reduction in the overall communication cost. For example, when the learning rate is set to 1000, the optimal cost is reached with 10 local communication rounds, making $K = 10$ a more efficient choice compared to a smaller number. On the other hand, at a lower learning rate of 100, the optimal cost of 12 is reached with $K = 3$. This pattern indicates that as we increase the number of local communication rounds, the total cost can be reduced, and the optimal number of local communication rounds tends to increase with higher learning rates.

Our key *contributions* are summarized as follows:

- We formulate *stochastic proximal point with arbitrary sampling* (SPPM-AS), a generalization of SPPM tailored to cross-device FL. SPPM-AS associates to any client-sampling distribution a pair of constants $(\mu_{\mathrm{AS}}, \sigma^2_{\star,\mathrm{AS}})$ that determines the convergence rate and neighborhood, and recovers FedAvg/FedProx-style schemes as special cases.

- We show that, in practice, the cohort proximal subproblem can be solved only approximately using K intra-cohort communication rounds of a standard FL optimizer $\mathcal{A}$ (LocalGD, CG, BFGS, Adam, etc.), and we analyze this inexact prox setting. This perspective explains how reusing the same cohort for multiple local rounds implements an SPPM-AS update and leads to substantial communication savings.

- We investigate several sampling strategies (NICE, block, stratified with clustering) and prove sampling-dependent bounds, including conditions under which stratified sampling yields a strictly smaller convergence neighborhood than NICE or block sampling. Experiments on logistic regression and CNNs in both standard and hierarchical FL setups show that appropriate choices of (sampling, $K, \mathcal{A}, \gamma$) reduce total communication cost by up to 74% (standard) and 95% (hierarchical) compared to FedAvg/LocalGD.

Objective. Our objective is to minimize total communication $C(\varepsilon, K)$ at a target accuracy ε. Theorem 2.4 expresses iteration progress via $(\mu_{\mathrm{AS}}, \sigma^2_{\star,\mathrm{AS}})$, determined by the client-sampling law $\mathcal{S}$ (Section 2.4). Together with the inexact-prox recursion (Lemma G.17), this links the number of local rounds K to the number of global rounds T. We formalize the resulting cost in a boxed corollary (after Theorem 2.4) and validate it empirically in Section 3.3 and 3.6 (Figs. 1–4).

2 Method

In this section, we explore efficient stochastic proximal point methods with arbitrary sampling for cross-device FL to optimize the objective equation 1. Throughout the paper, we denote $[n] := \{1, \ldots, n\}$. Our approach builds on the following assumptions.

Assumption 2.1. *Function $f_i : \mathbb{R}^d \to \mathbb{R}$ is differentiable for all samples $i \in [n]$.*

This implies that the function f is differentiable. The order of differentiation and summation can be interchanged due to the additive property of the gradient operator. $\nabla f(x) \overset{Eqn.\ (1)}{=} \nabla \left[\frac{1}{n}\sum_{i=1}^n f_i(x)\right] = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x)$.

Assumption 2.2. *Function $f_i : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex for all samples $i \in [n]$, where $\mu > 0$. That is, $f_i(y) + \langle \nabla f_i(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2 \le f_i(x)$, for all $x, y \in \mathbb{R}^d$.*

This implies that f is μ-strongly convex and hence has a unique minimizer, which we denote by $x_\star$. We know that $\nabla f(x_\star) = 0$. Notably, we do *not* assume f to be L-smooth.

2.1 Background: Proximal Point Methods in FL

We work with the standard cross-device FL objective in (1), where f_i is the empirical loss on client i. In a FedAvg-style round, the server samples a cohort C_t, broadcasts x_t to C_t, clients update locally on f_i, and their models are averaged.

The proximal point method (PPM) (Moreau, 1965) for minimizing a convex function φ generates

$$x_{t+1} = \mathrm{prox}_{\gamma\varphi}(x_t) := \arg\min_{z\in\mathbb{R}^d}\left\{\varphi(z) + \frac{1}{2\gamma}\|z - x_t\|^2\right\}, \qquad (2)$$

i.e., a regularized minimization of φ around x_t. Stochastic variants (SPPM) (Khaled & Jin, 2023) replace φ with a random function f_ξ drawn at each iteration.

Our proposed SPPM-AS instantiates this idea in cross-device FL: at iteration t we sample a cohort C_t from a distribution over client subsets, define its objective f_{C_t}, and perform one (possibly inexact) stochastic proximal step $x_{t+1} \approx \mathrm{prox}_{\gamma f_{C_t}}(x_t)$. The next subsections formalize this construction and introduce the sampling-dependent constants $(\mu_{\mathrm{AS}}, \sigma^2_{\star,\mathrm{AS}})$.

2.2 Sampling Distribution

Let $\mathcal{S}$ be a probability distribution over the 2^n subsets of $[n]$. Given a random set $S \sim \mathcal{S}$, we define

$$p_i := \mathrm{Prob}(i \in S), \quad i \in [n].$$

We restrict our attention to proper and nonvacuous random sets.

Assumption 2.3. *$\mathcal{S}$ is proper (i.e., $p_i > 0$ for all $i \in [n]$) and nonvacuous (i.e.,* $\mathrm{Prob}(S = \emptyset) = 0$*).*

Let C be the selected cohort. Given $\emptyset \neq C \subseteq [n]$ and $i \in [n]$, we define

$$v_i(C) := \begin{cases} \frac{1}{p_i} & i \in C \\ 0 & i \notin C \end{cases}, \quad f_C(x) \quad := \frac{1}{n}\sum_{i=1}^{n} v_i(C) f_i(x) = \sum_{i \in C} \frac{1}{np_i} f_i(x). \tag{3}$$

Note that $v_i(S)$ is a random variable and f_S is a random function. By construction, $\mathrm{E}_{S\sim\mathcal{S}}\left[v_i(S)\right] = 1$ for all $i \in [n]$, and hence

$$\mathrm{E}_{S\sim\mathcal{S}}\left[f_S(x)\right] = \mathrm{E}_{S\sim\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n} v_i(S) f_i(x)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_{S\sim\mathcal{S}}\left[v_i(S)\right] f_i(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) = f(x).$$

Therefore, the optimization problem in Equation (1) is equivalent to the stochastic optimization problem:

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \mathrm{E}_{S\sim\mathcal{S}}\left[f_S(x)\right] \right\}. \tag{4}$$

Further, if for each $C \subset [n]$ we let $p_C := \mathrm{Prob}(S = C)$, then f can be written in the equivalent form

$$f(x) = \mathbb{E}[[]\, S \sim \mathcal{S}] f_S(x) = \sum_{C \subseteq [n]} p_C f_C(x) = \sum_{C \subseteq [n], p_C > 0} p_C f_C(x). \tag{5}$$

Communication-cost model. We optimize communication via $C_{\text{flat}}(\varepsilon, K) = K\,T(\varepsilon)$ in the standard setting (Figs. 1–2) and $C_{\text{hier}}(\varepsilon, K) = (c_1 K + c_2)\,T(\varepsilon)$ in hierarchical FL (Sec. 3.6; Fig. 2d, 4). Both depend on the sampling law $\mathcal{S}$ only through $(\mu_{\text{AS}}, \sigma^2_{\star,\text{AS}})$ in Eq. (6), hence the sampling analysis directly informs communication.

2.3 Core Algorithm

For any proper, closed, convex function $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ and stepsize $\gamma > 0$ we define the (scaled) proximal operator $\mathrm{prox}_{\gamma\varphi}(x)$ as in (2). Intuitively, $\mathrm{prox}_{\gamma\varphi}(x)$ returns the point that compromises between minimizing φ and staying close to x.

Applying SPPM (Khaled & Jin, 2023) to Equation (4), we arrive at stochastic proximal point method with arbitrary sampling (SPPM-AS, Algorithm 1):

$$x_{t+1} = \mathrm{prox}_{\gamma f_{S_t}}(x_t),$$

where $S_t \sim \mathcal{S}$.

Algorithm 1 Stochastic Proximal Point Method with Arbitrary Sampling (SPPM-AS)

```
Input: starting point $x^0 \in \mathbb{R}^d$, distribution $\mathcal{S}$ over
    subsets of $[n]$, learning rate $\gamma > 0$
for $t = 0, 1, 2, \ldots$ do
    Sample $S_t \sim \mathcal{S}$
    $x_{t+1} = \mathrm{prox}_{\gamma f_{S_t}}(x_t)$
end for
```

Intuitive FL implementation. In our cross-device implementation, a single iteration of SPPM-AS proceeds as follows. The server samples a cohort C_t and broadcasts the model x_t to all clients in C_t. To approximate $x_{t+1} \approx \mathrm{prox}_{\gamma f_{C_t}}(x_t)$, the server then runs K *local communication rounds* with this fixed cohort: in each round k, clients in C_t receive the current model, perform several local optimization steps on their f_i, send updates back, and the server aggregates them into a new shared iterate. After K such rounds, the final cohort model $x_{t,K}$ is returned to the server as x_{t+1}. Thus, "reusing a cohort" means solving a single proximal subproblem more accurately by repeated server–cohort synchronizations, not keeping the same cohort across unrelated global iterations. We present the FL-instantiated algorithm in Appendix D.5.

Theorem 2.4 (Convergence of SPPM-AS)**.** *Let Assumption 2.1 (differentiability) and Assumption 2.2 (strong convexity) hold. Let $\mathcal{S}$ be a sampling satisfying Assumption 2.3, and define*

$$\mu_{\mathrm{AS}} := \min_{C \subseteq [n], p_C > 0} \sum_{i \in C} \frac{\mu_i}{n p_i}, \quad \sigma^2_{\star,\mathrm{AS}} := \sum_{C \subseteq [n], p_C > 0} p_C \left\| \nabla f_C \left(x_\star \right) \right\|^2 . \tag{6}$$

Let $x_0 \in \mathbb{R}^d$ be an arbitrary starting point. Then for any $t \geq 0$ and any $\gamma > 0$, the iterates of SPPM-AS *(Algorithm 1) satisfy*

$$\mathrm{E}\left[\|x_t - x_\star\|^2\right] \leq \left(\frac{1}{1 + \gamma \mu_{\mathrm{AS}}} \right)^{2t} \|x_0 - x_\star\|^2 + \frac{\gamma \sigma^2_{\star,\mathrm{AS}}}{\gamma \mu^2_{\mathrm{AS}} + 2\mu_{\mathrm{AS}}}.$$

Communication-cost objective. Let $T(\varepsilon)$ denote the number of SPPM-AS iterations required to reach $\mathbb{E}\|x_t - x^\star\|^2 \leq \varepsilon$ as given by Theorem 2.4 (see App. G.5 for the explicit expression). We measure total communication as

$$C_{\text{flat}}(\varepsilon, K) = K\,T(\varepsilon), \qquad C_{\text{hier}}(\varepsilon, K) = (c_1 K + c_2)\,T(\varepsilon),$$

corresponding respectively to flat and hierarchical FL. These are exactly the metrics reported in Figs. 1–2 (TK) and Figs. 2d, 4 ($(c_1K{+}c_2)T$). In App. G.10, Prop. G.18 shows that if the K-step local prox solver reduces its error geometrically, i.e., $\mathbb{E}|\operatorname{prox}^A_{\gamma f_S}(x) - \operatorname{prox}\gamma f_S(x)|^2 \leq B\rho^K$ with $0 < \rho < 1$, then the total communication cost $C_{\text{tot}}(\varepsilon; K)$ is minimized at some finite $K^\star$. Intuitively, increasing K makes each global round linearly more expensive (factor $(c_1K + c_2)$) but decreases the prox inexactness geometrically ($B\rho^K$), so their product has a unique sweet-spot $K^\star$.

Theorem interpretation. In Theorem 2.4, there are two main terms: $\left({}^1\!/\!(1+\gamma\mu_{\mathrm{AS}})\right)^{2t}$ and $\gamma\sigma^2_{\star,\mathrm{AS}}/(\gamma\mu^2_{\mathrm{AS}}+2\mu_{\mathrm{AS}})$, which define the convergence speed and neighborhood, respectively. Additionally, there are three hyperparameters to control the behavior: γ (the global learning rate), AS (the sampling type), and T (the number of global iterations). In the following paragraphs, we will explore special cases to provide a clear intuition of how the SPPM-AS theory works.

Interpolation regime. Consider the interpolation regime, characterized by $\sigma^2_{\star,\mathrm{AS}} = 0$. Since we can use arbitrarily large $\gamma > 0$, we obtain an arbitrarily fast convergence rate. Indeed, $\left({}^1\!/\!(1+\gamma\mu_{\mathrm{AS}})\right)^{2t}$ can be made arbitrarily small for any fixed $t \geq 1$, even $t = 1$, by choosing γ large enough. However, this is not surprising, since now f and all functions f_ξ share a single minimizer, $x_\star$, and hence it is possible to find it by sampling a small batch of functions even a single function f_ξ, and minimizing it, which is what the prox does, as long as γ is large enough.

A single step travels far. Observe that for $\gamma = {}^1\!/\mu_{\mathrm{AS}}$, we have $\gamma\sigma^2_{\star,\mathrm{AS}}/(\gamma\mu^2_{\mathrm{AS}}+2\mu_{\mathrm{AS}}) = \sigma^2_{\star,\mathrm{AS}}/3\mu^2_{\mathrm{AS}}$. In fact, the convergence neighborhood $\gamma\sigma^2_{\star,\mathrm{AS}}/(\gamma\mu^2_{\mathrm{AS}}+2\mu_{\mathrm{AS}})$ is bounded above by three times this quantity irrespective of the choice of the stepsize. Indeed, $\frac{\gamma\sigma^2_{\star,\mathrm{AS}}}{\gamma\mu^2_{\mathrm{AS}}+2\mu_{\mathrm{AS}}} \leq \min\left\{\frac{\sigma^2_{\star,\mathrm{AS}}}{\mu^2_{\mathrm{AS}}}, \frac{\gamma\sigma^2_{\star,\mathrm{AS}}}{\mu_{\mathrm{AS}}}\right\} \leq \frac{\sigma^2_{\star,\mathrm{AS}}}{\mu^2_{\mathrm{AS}}}$. That means that no matter how far the starting point x_0 is from the optimal solution $x_\star$, if we choose the stepsize γ to be large enough, then we can get a decent-quality solution after a single iteration of SPPM-AS already! Indeed, if we choose γ large enough so that $\left({}^1\!/1+\gamma\mu_{\mathrm{AS}}\right)^2 \left\|x_0 - x_\star\right\|^2 \leq \delta$, where $\delta > 0$ is chosen arbitrarily, then for $t = 1$ we get $\mathbb{E}\left[\left\|x_1 - x_\star\right\|^2\right] \leq \delta + \frac{\sigma^2_{\star,\mathrm{AS}}}{\mu^2_{\mathrm{AS}}}$.

Iteration complexity. We have seen above that an accuracy arbitrarily close to (but not reaching) $\sigma^2_{\star,\mathrm{AS}}/\mu^2_{\mathrm{AS}}$ can be achieved via a single step of the method, provided that the stepsize γ is large enough. Assume now that we aim for ϵ accuracy, where $\epsilon \leq \sigma^2_{\star,\mathrm{AS}}/\mu^2_{\mathrm{AS}}$. We can show that with the stepsize $\gamma = \varepsilon\mu_{\mathrm{AS}}/\sigma^2_{\star,\mathrm{AS}}$, we get $\mathrm{E}\left[\left\|x_t - x_\star\right\|^2\right] \leq \varepsilon$ provided that $t \geq \left(\frac{\sigma^2_{\star,\mathrm{AS}}}{2\varepsilon\mu^2_{\mathrm{AS}}} + \frac{1}{2}\right)\log\left(\frac{2\|x_0 - x_\star\|^2}{\varepsilon}\right)$. We provide the proof in Appendix G.5. To ensure thoroughness, we present in Appendix G.9 the lemma of the inexact formulation for SPPM-AS, which offers greater practicality for empirical experimentation. Further insights are provided in the subsequent experimental section.

General framework. With freedom to choose arbitrary algorithms for solving the proximal operator one can see that SPPM-AS is generalization for such renowned methods as FedProx (Li et al., 2020a) and FedAvg (McMahan et al., 2016). A more particular overview of FedProx-SPPM-AS is presented in further Appendix C.4.

2.4 Arbitrary Sampling Examples

Why sampling matters for communication. In SPPM-AS, the sampling law $\mathcal{S}$ impacts communication only through the rate constants $(\mu_{\mathrm{AS}}, \sigma^2_{\star,\mathrm{AS}})$ in Eq. (6), which determine the iteration count $T(\varepsilon)$ appearing in $C(\varepsilon, K)$. Thus, the sampling analysis in this section is not ancillary: it is the mechanism by which communication cost is predicted and reduced.

Details on simple Full Sampling (FS) and Nonuniform Sampling (NS) are provided in Appendix C.2. In this section, we focus more intently on the sampling strategies that are of particular interest to us.

Nice Sampling (NICE). Choose $\tau \in [n]$ and let S be a random subset of $[n]$ of size τ chosen uniformly at random. Then $p_i = {}^{\tau}\!/_{n}$ for all $i \in [n]$. Moreover, let $\binom{n}{\tau}$ represents the number of combinations of n taken τ at a time, $p_C = \frac{1}{\binom{n}{\tau}}$ whenever $|C| = \tau$ and $p_C = 0$ otherwise. So,

$$\mu_{\mathrm{AS}} = \mu_{\mathrm{NICE}}(\tau) := \min_{C \subseteq [n], p_C > 0} \sum_{i \in C} \frac{\mu_i}{np_i} = \min_{C \subseteq [n], |C| = \tau} \frac{1}{\tau} \sum_{i \in C} \mu_i,$$

$$\sigma^2_{\star,\mathrm{AS}} = \sigma^2_{\star,\mathrm{NICE}}(\tau) := \sum_{C \subseteq [n], p_C > 0} p_C \left\| \nabla f_C\left(x_\star\right) \right\|^2 \overset{Eqn.~(3)}{=} \sum_{C \subseteq [n], |C| = \tau} \frac{1}{\binom{n}{\tau}} \left\| \frac{1}{\tau} \sum_{i \in C} \nabla f_i\left(x_\star\right) \right\|^2.$$

It can be shown that $\mu_{\mathrm{NICE}}(\tau)$ is a *nondecreasing* function of τ (Appendix G.6). So, as the minibatch size τ increases, the strong convexity constant $\mu_{\mathrm{NICE}}(\tau)$ can only improve. Since $\mu_{\mathrm{NICE}}(1) = \min_i \mu_i$ and $\mu_{\mathrm{NICE}}(n) = \frac{1}{n}\sum_{i=1}^n \mu_i$, the value of $\mu_{\mathrm{NICE}}(\tau)$ interpolates these two extreme cases as τ varies between 1 and n. Conversely, $\sigma^2_{\star,\mathrm{NICE}}(\tau) = \frac{n/\tau - 1}{n-1}\sigma^2_{\star,\mathrm{NICE}}(1)$ is a nonincreasing function, reaching a value of $\sigma^2_{\star,\mathrm{NICE}}(n) = 0$, as explained in Appendix G.6.

Block Sampling (BS). Let $C_1, \ldots, C_b$ be a partition of $[n]$ into b nonempty blocks. For each $i \in [n]$, let $B(i)$ indicate which block i belongs to. In other words, $i \in C_j$ if $B(i) = j$. Let $S = C_j$ with probability $q_j > 0$, where $\sum_j q_j = 1$. Then $p_i = q_{B(i)}$, and hence Equation (6) takes on the form

$$\mu_{\mathrm{AS}} = \mu_{\mathrm{BS}} := \min_{j \in [b]} \frac{1}{nq_j} \sum_{i \in C_j} \mu_i, \quad \sigma^2_{\star,\mathrm{AS}} = \sigma^2_{\star,\mathrm{BS}} := \sum_{j \in [b]} q_j \left\| \sum_{i \in C_j} \frac{1}{np_i} \nabla f_i\left(x_\star\right) \right\|^2.$$

Considering two extreme cases: If $b = 1$, then SPPM-BS = SPPM-FS = PPM. So, indeed, we recover the same rate as SPPM-FS. If $b = n$, then SPPM-BS = SPPM-NS. So, indeed, we recover the same rate as SPPM-NS. We provide the detailed analysis in Appendix C.3.

Stratified Sampling (SS). Let $C_1, \ldots, C_b$ be a partition of $[n]$ into b nonempty blocks, as before. For each $i \in [n]$, let $B(i)$ indicate which block does i belong to. In other words, $i \in C_j$ iff $B(i) = j$. Now, for each $j \in [b]$ pick $\xi_j \in C_j$ uniformly at random, and define $S = \cup_{j \in [b]} \{\xi_j\}$. Clearly, $p_i = \frac{1}{|C_{B(i)}|}$. Let's denote $\mathbf{i}_b := (i_1, \cdots, i_b)$, $\mathbf{C}_b := C_1 \times \cdots \times C_b$. Then, Equation (6) take on the form

$$\mu_{\mathrm{AS}} = \mu_{\mathrm{SS}} := \min_{\mathbf{i}_b \in \mathbf{C}_b} \sum_{j=1}^b \frac{\mu_{i_j} |C_j|}{n}, \quad \sigma^2_{\star,\mathrm{AS}} = \sigma^2_{\star,\mathrm{SS}} := \sum_{\mathbf{i}_b \in \mathbf{C}_b} \left(\prod_{j=1}^b \frac{1}{|C_j|} \right) \left\| \sum_{j=1}^b \frac{|C_j|}{n} \nabla f_{i_j}\left(x_\star\right) \right\|^2.$$

Lemma 2.5 (Stratified Sampling Variance Bounds)**.** *Consider the stratified sampling. For each $j \in [b]$, define $\sigma_j^2 := \max_{i \in C_j} \left\| \nabla f_i\left(x_\star\right) - \frac{1}{|C_j|} \sum_{l \in C_j} \nabla f_l\left(x_\star\right) \right\|^2$. In words, σ_j^2 is the maximal squared distance of a gradient (at the optimum) from the mean of the gradients (at optimum) within cluster C_j. Then $\sigma^2_{\star,\mathrm{SS}} \leq \frac{b}{n^2} \sum_{j=1}^b |C_j|^2 \sigma_j^2 \leq b \max\left\{\sigma_1^2, \ldots, \sigma_b^2\right\}$.*

Considering two extreme cases: If $b = 1$, then SPPM-SS = SPPM-US. So, indeed, we recover the same rate as SPPM-US. If $b = n$, then SPPM-SS = SPPM-FS. So, indeed, we recover the same rate as SPPM-FS. We provide the detailed analysis in Appendix C.3.

Note that Lemma 2.5 provides insights into how the variance might be reduced through stratified sampling. For instance, in a scenario of complete inter-cluster homogeneity, where $\sigma_j^2 = 0$ for all j, both bounds imply that $0 = \sigma^2_{\star,\mathrm{SS}} \le \sigma^2_{\star,\mathrm{BS}}$. Thus, in this scenario, the convergence neighborhood of stratified sampling is better than that of block sampling.

Stratified sampling outperforms block sampling and nice sampling in convergence neighborhood. We theoretically compare stratified sampling with block sampling and nice sampling, advocating for stratified sampling as the superior method for future clustering experiments due to its optimal variance properties. We begin with the assumption of b clusters of uniform size b (Assumption G.12), which simplifies the analysis by enabling comparisons of various sampling methods, all with the same sampling size, b: b-nice sampling, stratified sampling with b clusters, and block sampling where all clusters are of uniform size b. Furthermore, we introduce the concept of optimal clustering for stratified sampling (noted as $\mathcal{C}_{b,\mathrm{SS}}$, Definition G.14) in response to a counterexample where block sampling and nice sampling achieve lower variance than stratified sampling (Example G.13). Finally, we compare neighborhoods using the stated assumption.

Lemma 2.6. *Given Assumption G.12, the following holds: $\sigma^2_{\star,\mathrm{SS}}\left(\mathcal{C}_{b,\mathrm{SS}}\right) \le \sigma^2_{\star,\mathrm{NICE}}$ for arbitrary b. Moreover, the variance within the convergence neighborhood of stratified sampling is less than or equal to that of nice sampling: $\frac{\gamma\sigma^2_{\star,\mathrm{SS}}}{\gamma\mu^2_{\mathrm{SS}}+2\mu_{\mathrm{SS}}}\left(\mathcal{C}_{b,\mathrm{SS}}\right) \le \frac{\gamma\sigma^2_{\star,\mathrm{NICE}}}{\gamma\mu^2_{\mathrm{NICE}}+2\mu_{\mathrm{NICE}}}$.*

Lemma 2.6 demonstrates that, under specific conditions, the stratified sampling neighborhood is preferable to that of nice sampling. One might assume that, under the same assumptions, a similar assertion could be made for showing that block sampling is inferior to stratified sampling . However, this has only been verified for the simplified case where both the block size and the number of blocks are $b = 2$, as detailed in Appendix G.8.

3 Experiments

Practical decision-making with SPPM-AS. In our analysis of SPPM-AS, guided by the theoretical foundations of Theorem 2.4 and empirical evidence summarized in Table 1, we explore practical decision-making for varying scenarios. This includes adjustments in hyperparameters within the framework $KT(\epsilon, \mathcal{S}, \gamma, \mathcal{A}(K))$. Here, ϵ represents accuracy goal, $\mathcal{S}$ the sampling distribution, γ the global learning rate (proximal operator parameter), $\mathcal{A}$ the proximal optimizer, and K the number of local communication rounds.

Table 1: $KT(\epsilon, \mathcal{S}, \gamma, \mathcal{A}(K))$

HP	Control	$KT(\cdots)$	Exp.
γ	$\gamma \uparrow$	$KT \downarrow$, $\epsilon \uparrow$ [1]	E.2
	optimal $(\gamma, K) \uparrow$	$\downarrow$	3.3
$\mathcal{A}$	μ-convex + BFGS/CG	$\downarrow$ vs. LocalGD	3.3
	NonCVX + Hierarchical FL + Adam	$\downarrow$ vs. LocalGD	3.7

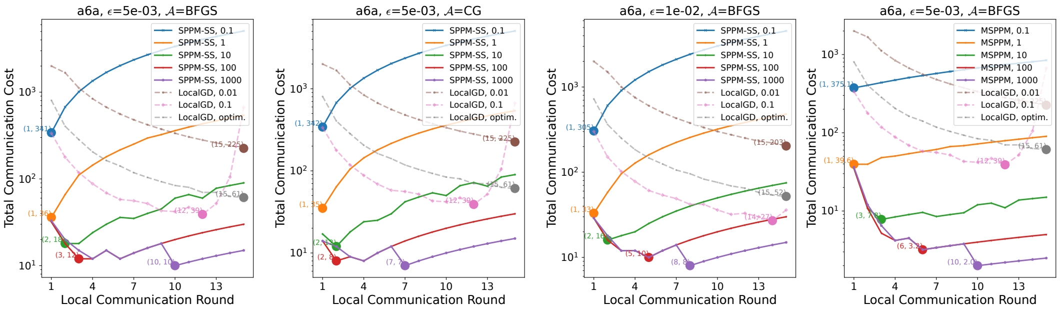
[1] ϵ is convergence neighborhood or accuracy.

In Table 1, we summarize how changes in these hyperparameters influence the target metric. Increasing γ leads to faster convergence but lower accuracy, demonstrating an accuracy-speed tradeoff. Our primary observation is that jointly increasing both γ and K improves the convergence rate. Moreover, employing different proximal solvers yields better results than FedAvg in both convex and non-convex cases.

3.1 Objective and Datasets

Our analysis begins with logistic regression with an l_2 regularizer, which can be represented as:

$$f_i(x) := \frac{1}{n_i}\sum_{j=1}^{n_i} \log\left(1 + \exp(-b_{i,j}x^T a_{i,j})\right) + \frac{\mu}{2}\|x\|^2,$$



(a) base (b) different prox solver (c) varying ϵ (d) hierarchical: $c_1 = 0.1$

Figure 2: Analysis of total communication costs against local communication rounds for computing the proximal operator. For LocalGD, we align the x-axis to the total local iterations, highlighting the absence of local communication. The aim is to minimize total communication for achieving a predefined global accuracy ϵ, where $\|x_T - x_\star\|^2 < \epsilon$. The optimal step size and minibatch sampling setup for LocalGD are denoted as LocalGD, optim. This showcases a comparison across varying ϵ values and proximal operator solvers (CG and BFGS).

where μ is the regularization parameter, n_i denotes the total number of data points at client i, $a_{i,j}$ are the feature vectors, and $b_{i,j} \in \{-1, 1\}$ are the corresponding labels. Each function f_i exhibits μ-strong convexity and L_i-smoothness, with L_i computed as $\frac{1}{4n_i}\sum_{j=1}^{n_i}\|a_{i,j}\|^2 + \mu$. For our experiments, we set μ to 0.1.

Our study utilized datasets from the LibSVM repository (Chang & Lin, 2011), including `mushrooms`, `a6a`, `ijcnn1.bz2`, and `a9a`. We divided these into feature-wise heterogeneous non-iid splits for FL, detailed in Appendix D.3, with a default cohort size of 10. We primarily examined logistic regression, finding results consistent with our theoretical framework, as discussed extensively in Section 3.3 through Appendix E.2. Additional neural network experiments are detailed in Section 3.7 and Appendix F.

3.2 On Choosing Sampling Strategy

As shown in Section 2.4, multiple sampling techniques exist. We propose using clustering approach in conjuction with SPPM-SS as the default sampling strategy for all our experiments. The stratified sampling optimal clustering is impractical due to the difficulty in finding $x_\star$; therefore, we employ a clustering heuristic that aligns with the concept of creating homogeneous worker groups. One such method is K-means, which we use by default. More details on our clustering approach can be found in the Appendix D.3. We compare various sampling techniques in the left panel of Figure 3. Extensive ablations verified the efficiency of stratified sampling over other strategies, due to variance reduction (Lemma 2.5).

3.3 Communication Reduction via Increased Local Rounds

In this study, we investigate whether increasing the number of local communication rounds, denoted as K, in our proposed algorithm SPPM-SS, can lead to a decrease in the total communication cost required to converge to a predetermined global accuracy $\epsilon > 0$. In Figure 1, we analyzed various datasets, including `a6a` and `mushrooms`, confirming that higher local communication rounds reduce communication costs, especially with larger learning rates. Our study includes both self-ablation of SPPM-SS across different learning rate scales and comparisons with the widely-used cross-device FL method LocalGD (or FedAvg) on the selected cohort. Ablation studies were conducted with a large empirical learning rate of 0.1, a smaller rate of 0.01, and an optimal rate as discussed by Khaled & Richtárik (2023), alongside minibatch sampling described by Gower et al. (2019).

In Figure 2, we present more extensive ablations. Specifically, we set the `base` method (Figure 2a) using the dataset a6a, a proximal solver BFGS, and $\epsilon = 5 \cdot 10^{-3}$. In Figure 2b, we explore the use of an alternative solver, CG (Conjugate Gradient), noting some differences in outcomes. For instance, with a learning rate $\gamma = 1000$, the optimal K with CG becomes 7, lower than 10 in the `base` setting

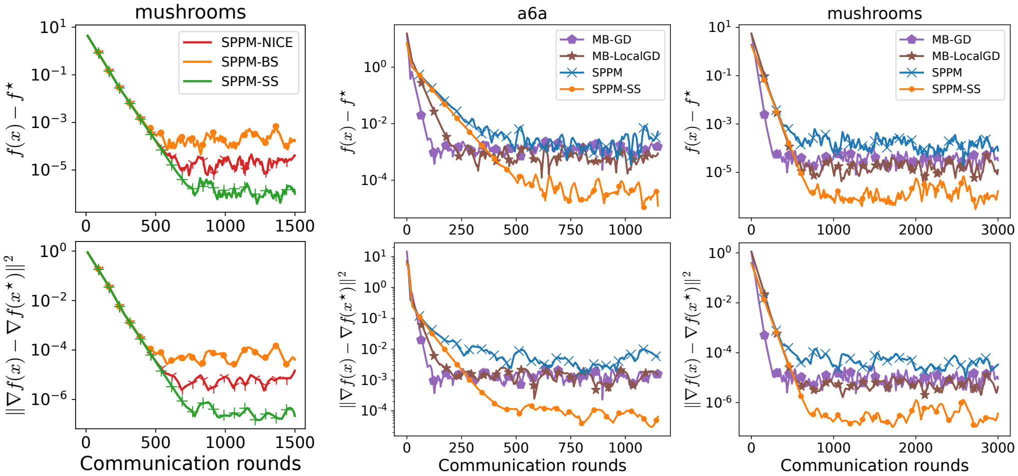


Figure 3: The first column compares sampling methods, while the right two columns analyze convergence relative to popular baselines. $\gamma = 1.0$.

using BFGS. In Figure 2c, we investigate the impact of varying $\epsilon = 10^{-2}$. Our findings consistently show SPPM-SS's significant performance superiority over LocalGD.

3.4 Evaluating the Performance of Various Solvers $\mathcal{A}$

We further explore the impact of various solvers on optimizing the proximal operators, showcasing representative methods in Table 2 in the Appendix B.3. A detailed overview and comparison of local optimizers listed in the table are provided in Section B.3, given the extensive range of candidate options available. To emphasize key factors, we compare the performance of first-order methods, such as the Conjugate Gradient (CG) method (Hestenes et al., 1952), against second-order methods, like the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm (Broyden, 1967; Shanno, 1970), in the context of strongly convex settings. For non-convex settings, where first-order methods are prevalent in deep learning experiments, we examine an ablation among popular first-order local solvers, specifically choosing MimeLite (Karimireddy et al., 2020a) and FedOpt (Reddi et al., 2020). The comparisons of different solvers for strongly convex settings are presented in Figure 2b, with the non-convex comparison included in the appendix. Upon comparing first-order and second-order solvers in strongly convex settings, we observed that CG outperforms BFGS for our specific problem. In neural network experiments, MimeLite-Adam was found to be more effective than FedOpt variations. However, it is important to note that all these solvers are viable options that have led to impressive performance outcomes.

3.5 Comparative Analysis with Baseline Algorithms

In this section, we conduct an extensive comparison with several established cross-device FL baseline algorithms. Specifically, we examine MB-GD (MiniBatch Gradient Descent with partial client participation), and MB-LocalGD, which is the local gradient descent variant of MB-GD. We default the number of local iterations to 5 and adopt the optimal learning rate as suggested by Gower et al. (2019). To ensure a fair comparison, the cohort size $|C|$ is fixed at 10 for all minibatch methods, including our proposed SPPM-SS. The results of this comparative analysis are depicted in Figure 3. Our findings reveal that SPPM-SS consistently achieves convergence within a significantly smaller neighborhood when compared to the existing baselines. Notably, in contrast to MB-GD and MB-LocalGD, SPPM-SS is capable of utilizing arbitrarily large learning rates. This attribute allows for faster convergence, although it does result in a larger neighborhood size.

3.6 Hierarchical Federated Learning

We extend our analysis to a hub-based hierarchical FL structure, as conceptualized in the left part of Figure 4. This structure envisions a cluster directly connected to m hubs, with each hub m_i serving n_i clients. The clients, grouped based on criteria such as region, communicate exclusively with their respective regional hub, which in turn communicates with the central server. Given the inherent na-

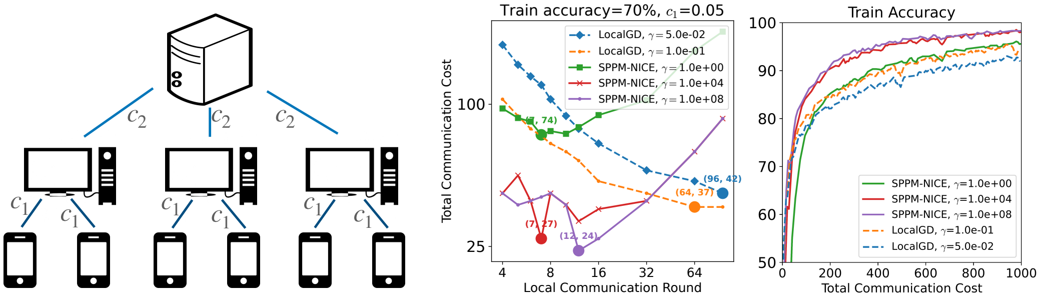


Figure 4: The left column shows the Server-hub-client hierarchical FL architecture. For the right two columns: on the left, communication cost for achieving 70% accuracy in hierarchical FL ($c_1 = 0.05$, $c_2 = 1$); on the right, convergence with optimal hyperparameters ($c_1 = 0.05$, $c_2 = 1$). In the hierarchical setting we report $(c_1 K + c_2)T$, matching the cost definition in Sec. 2.3; here $c_1 \ll c_2$.

ture of this hierarchical model, the communication cost c_1 from each client to its hub is consistently lower than the cost c_2 from each hub to the server. We define communication from clients to hubs as *local communication* and from hubs to the server as *global communication*. Under SPPM-SS, the total cost is expressed as $(c_1 K + c_2)T_{\text{SPPM-SS}}$, while for LocalGD, it is $(c_1 + c_2)T_{\text{LocalGD}}$. As established in Section 3.3, $T_{\text{SPPM-SS}}$ demonstrates significant improvement in total communication costs compared to LocalGD within a hierarchical setting. Our objective is to illustrate this by contrasting the standard FL setting, depicted in Figure 2a with parameters $c_1 = 1$ and $c_2 = 0$, against the hierarchical FL structure, which assumes $c_1 = 0.1$ and $c_2 = 1$, as shown in Figure 2d. Given the variation in c_1 and c_2 values between these settings, a direct comparison of absolute communication costs is impractical. Therefore, our analysis focuses on the ratio of communication cost reduction in comparison to LocalGD. For the `base` setting, LocalGD's optimal total communication cost is 39 with 12 local iterations, whereas for SPPM-SS ($\gamma = 1000$), it is reduced to 10 with 10 local and 1 global communication rounds, amounting to a 74.36% reduction. With the hierarchical FL structure in Figure 2d, SPPM-SS achieves an even more remarkable communication cost reduction of 94.87%. Further ablation studies on varying local communication cost c_1 in the Appendix E.3 corroborate these findings.

3.7 Neural Network Evaluations

Our empirical analysis includes experiments on Convolutional Neural Networks (CNNs) using the `FEMNIST` dataset, as described by Caldas et al. (2018). We designed the experiments to include a total of 100 clients, with each client representing data from a unique user, thereby introducing natural heterogeneity into our study. We employed the Nice sampling strategy with a cohort size of 10. In contrast to logistic regression models, here we utilize training accuracy as a surrogate for the target accuracy ϵ. For the optimization of the proximal operator, we selected the Adam optimizer, with the learning rate meticulously fine-tuned over a linear grid. Detailed descriptions of the training procedures and the CNN architecture are provided in the Appendix F.

Our analysis primarily focuses on the hierarchical FL structure. Initially, we draw a comparison between our proposed method, SPPM-AS, and LocalGD. The crux of our investigation is the total communication cost required to achieve a predetermined level of accuracy, with findings detailed in the right part of Figure 4. Significantly, SPPM-AS demonstrates enhanced performance with the integration of multiple local communication rounds. Notably, the optimal number of these rounds tends to increase alongside the parameter γ. For each configuration, the convergence patterns corresponding to the sets of optimally tuned hyperparameters are depicted in Figure 4.

4 Conclusion

Our work revisits the standard *one communication round per cohort* design in cross-device federated learning. By formulating cohort reuse as a stochastic proximal point method with arbitrary sampling (SPPM-AS), we show theoretically and empirically that allowing $K > 1$ intra-cohort rounds can substantially reduce total communication cost, up to 74% compared with FedAvg-style baselines on both convex and non-convex tasks. We hope this cohort-squeeze perspective inspires more communication-aware federated learning algorithms and systems.

References

Mehdi Salehi Heydar Abad, Emre Ozfatura, Deniz Gunduz, and Ozgur Ercetin. Hierarchical federated learning across heterogeneous cellular networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8866–8870. IEEE, 2020.

Hilal Asi and John C Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019.

Hilal Asi, Karan Chadha, Gary Cheng, and John C Duchi. Minibatch stochastic approximate proximal point methods. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21958–21968. Curran Associates, Inc., 2020.

Dimitri P Bertsekas. Incremental proximal methods for large scale convex optimization. *Mathematical Programming*, 129(2):163–195, 2011.

Sebastian Bischoff, Stephan Günnemann, Martin Jaggi, and Sebastian U. Stich. On second-order optimization methods for federated learning. *arXiv preprint arXiv:2303.10581*, 2023.

Charles G Broyden. Quasi-Newton methods and their application to function minimisation. *Mathematics of Computation*, 21(99):368–381, 1967.

Aysegul Bumin and Kejun Huang. Efficient implementation of stochastic proximal point algorithm for matrix and tensor completion. In *29th European Signal Processing Conference (EUSIPCO)*, pp. 1050–1054. IEEE, 2021.

Sebastian Caldas, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A benchmark for federated settings. 2018.

Karan Chadha, Gary Cheng, and John Duchi. Accelerated, optimal and parallel: Some results on model-based stochastic optimization. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 2811–2827. PMLR, 2022.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322, 1970.

Gerald B. Folland. Real Analysis: Modern Techniques and Their Applications. 1984.

Caroline Geiersbach and Teresa Scarinci. Stochastic proximal gradient methods for nonconvex problems in hilbert spaces. *Computational Optimization and Applications*, 78(3):705–740, 2021. doi: 10.1007/s10589-[]020-[]00259-[]y.

Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5200–5209. PMLR, 2019.

Rami Hamdi, Ahmed Ben Said, Aiman Erbad, Amr Mohamed, Mounir Hamdi, and Mohsen Guizani. Hierarchical federated learning over hetnets enabled by wireless energy transfer. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pp. 01–06. IEEE, 2021.

Slavomír Hanzely, Dmitry Kamzolov, Dmitry Pasechnyuk, Alexander Gasnikov, Peter Richtárik, and Martin Takáč. A damped Newton method achieves global $O\left(\frac{1}{k^2}\right)$ and local quadratic convergence rate, 2022.

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

Magnus Rudolph Hestenes, Eduard Stiefel, et al. *Methods of conjugate gradients for solving linear systems*, volume 49. NBS Washington, DC, 1952.

Majid Jahani, Sergey Rusakov, Zheng Shi, Peter Richtárik, Michael W Mahoney, and Martin Takáč. Doubly adaptive scaled algorithm for machine learning using second-order information. *arXiv preprint arXiv:2109.05198*, 2021.

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. FedExP: Speeding up federated averaging via extrapolation. *arXiv preprint arXiv:2301.09604*, 2023.

P. Kairouz et al. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2):1–210, 2019.

Belhal Karimi, Ping Li, and Xiaoyun Li. Layer-wise and dimension-wise locally adaptive federated learning, 2022.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pp. 795–811, 2016.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020a.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 5132–5143. PMLR, 2020b.

A. Khaled, K. Mishchenko, and P. Richtárik. First analysis of local GD on heterogeneous data. paper arXiv:1909.04715, presented at NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality, 2019.

Ahmed Khaled and Chi Jin. Faster federated optimization under second-order similarity. In *The Eleventh International Conference on Learning Representations*, 2023.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. 2020a.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-IID data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020b.

Dachao Lin, Yuze Han, Haishan Ye, and Zhihua Zhang. Stochastic distributed optimization under average second-order similarity: Algorithms and analysis. *Advances in Neural Information Processing Systems*, 36, 2024.

Xuezhe Ma. Apollo: An adaptive parameter-wise diagonal quasi-Newton method for nonconvex stochastic optimization. *arXiv preprint arXiv:2009.13586*, 2020.

Grigory Malinovsky, Konstantin Mishchenko, and Peter Richtárik. Server-side stepsizes and sampling without replacement provably help in federated optimization. In *Proceedings of the 4th International Workshop on Distributed Machine Learning*, pp. 85–104, 2023.

Grigory Malinovsky, Kai Yi, and Peter Richtárik. Variance reduced ProxSkip: Algorithm, theory and application to federated learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2024. ISBN 9781713871088.

Bernard Martinet. Regularisation d'inequations variationelles par approximations successives. *Revue Francaise d'informatique et de Recherche operationelle*, 4:154–159, 1970.

B. McMahan, E. Moore, D. Ramage, and B. Agüera y Arcas. Federated learning of deep networks using model averaging. 2016.

H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Konstantin Mishchenko, Ahmed Khaled, and Peter Richtarik. Proximal and federated random reshuffling. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 15718–15749. PMLR, 2022a.

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtarik. ProxSkip: Yes! Local gradient steps provably lead to communication acceleration! Finally! In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 15750–15769. PMLR, 2022b.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

Andrei Patrascu and Ion Necoara. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *Journal of Machine Learning Research*, 18(198):1–42, 2018.

Swaroop Ramaswamy, Rajiv Mathews, Kanishka Rao, and Françoise Beaufays. Federated learning for emoji prediction in a mobile keyboard. *arXiv preprint arXiv:1906.04329*, 2019.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

Ernest Ryu and Stephen Boyd. Stochastic proximal iteration: A non-asymptotic improvement upon stochastic gradient descent. Technical report, Stanford University, 2016.

David F Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970.

Alex Shtoff. Efficient implementation of incremental proximal-point methods. *arXiv preprint arXiv:2205.01457*, 2022.

Jianhui Sun, Xidong Wu, Heng Huang, and Aidong Zhang. On the role of server momentum in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15164–15172, 2024.

Yan Sun, Li Shen, Tiansheng Huang, Liang Ding, and Dacheng Tao. Fedspeed: Larger local interval, less communication round, and higher generalization accuracy. *arXiv preprint arXiv:2302.10429*, 2023.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. A novel framework for the analysis and design of heterogeneous federated learning. *IEEE Transactions on Signal Processing*, 69:5234–5249, 2021a.

Jianyu Wang, Zheng Xu, Zachary Garrett, Zachary Charles, Luyang Liu, and Gauri Joshi. Local adaptivity in federated learning: Convergence and consistency. *arXiv preprint arXiv:2106.02305*, 2021b.

Jing Xu, Sen Wang, Liwei Wang, and Andrew Chi-Chih Yao. FedCM: Federated learning with client-level momentum. *arXiv preprint arXiv:2106.10874*, 2021.

He Yang. H-fl: A hierarchical communication-efficient and privacy-protected architecture for federated learning. *IJCAI*, 2021.

Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving Google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*, 2018.

Kai Yi, Nidham Gazagnadou, Peter Richtárik, and Lingjuan Lyu. FedP3: Federated personalized and privacy-friendly network pruning under model heterogeneity. In *The Twelfth International Conference on Learning Representations*, 2024.

Kai Yi, Laurent Condat, and Peter Richtárik. Explicit personalization and local training: Double communication acceleration in federated learning. *Transactions on Machine Learning Research (TMLR)*, 2025.

Xiao-Tong Yuan and Ping Li. Sharper analysis for minibatch stochastic proximal point methods: Stability, smoothness, and deviation. *Journal of Machine Learning Research*, 24(270):1–52, 2023.

Xiaotong Yuan and Ping Li. On convergence of FedProx: Local dissimilarity invariant bounds, non-smoothness and beyond. *Advances in Neural Information Processing Systems*, 35:10752–10765, 2022.

Dun Zeng, Siqi Liang, Xiangjing Hu, Hui Wang, and Zenglin Xu. FedLab: A flexible federated learning framework. *Journal of Machine Learning Research*, 24(100):1–7, 2023.

CONTENTS

A Discussion, Limitations, and Future Work

This work introduces *Cohort Squeeze*, a novel framework that extends classical cross-device federated learning (FL) protocols by enabling multiple local communication rounds within a single cohort. Through a principled reformulation using the stochastic proximal point method with arbitrary sampling (SPPM-AS), we demonstrate both theoretically and empirically that increasing intra-cohort communication rounds can significantly reduce total communication cost—achieving up to 74% improvement over FedAvg and related baselines.

Feasibility of cohort reuse. Our analysis assumes that, once a cohort C_t is sampled at iteration t, the participating clients remain available for K intra-cohort communication rounds. This is analogous to standard cross-device FL, where clients selected for a round stay connected long enough to perform several local epochs before uploading. We do not require the same cohort across different global iterations. In practice, client dropouts during these K rounds can be handled by simply shrinking the effective cohort, which is captured by our arbitrary-sampling abstraction. More sophisticated replacement strategies are an interesting direction for future work.

While our results consistently support the benefits of cohort reuse and local communication amplification, several aspects merit further investigation. First, although we provide a rigorous convergence analysis under strong convexity assumptions, extending these results to more general non-convex objectives remains an open theoretical direction. Second, our empirical studies employ clustering-based stratified sampling, with default heuristics such as k-means. Although effective in our setting, these strategies may not fully exploit underlying client similarity in more heterogeneous or dynamically evolving environments. Third, while we examine representative solvers such as CG, BFGS, and Adam, further tuning or combining proximal solvers could yield additional gains in efficiency and generalization.

Looking ahead, a promising avenue lies in optimizing the joint schedule of global and local updates, possibly adapting the number of local rounds dynamically based on cohort statistics or server feedback. Moreover, the integration of privacy-preserving mechanisms (e.g., differential privacy or secure aggregation) into the multi-round cohort setting remains unexplored and is critical for practical deployment. Lastly, real-world deployment in edge computing and mobile scenarios would help validate the scalability and robustness of the proposed method under diverse network and resource constraints.

Overall, this work opens a new perspective in cross-device FL by demonstrating the untapped potential of each selected cohort, and we hope it stimulates future research into more adaptive, communication-efficient federated optimization paradigms.

Beyond strong convexity. Two proof points use strong convexity critically in our analysis: (i) the *contractivity of the prox* in Fact G.4, which yields the linear factor $(1+\gamma\mu_{\mathrm{AS}})^{-2}$ in the distance recursion, and (ii) the iteration bound in App. G.5, which instantiates this linear recurrence to obtain $T(\varepsilon)$. In App. G.11 we show how the same proof skeleton extends to two standard relaxations. First, under Polyak–Łojasiewicz (PL) or quadratic-growth (QG) conditions for f_C (or f), the prox remains strictly contractive around the minimizer set, and the SPPM-AS iteration enjoys linear convergence to a noise floor with μ_{AS} replaced by a PL/QG constant. Second, for weakly-convex (possibly non-convex) objectives we reinterpret SPPM-AS as a stochastic gradient method on the Moreau envelope of f and obtain sublinear convergence of a stationarity measure, with the inexact-prox Lemma G.17 translating the number of local rounds K into a bias term $b(K)$. Section 3.7 empirically fits this weakly-convex regime: the CNN experiments on FEMNIST use Adam to implement the inexact prox and exhibit the expected behavior as K and γ vary.

Broader Impact

This work proposes a novel method for improving communication efficiency in cross-device FL, a setting highly relevant for large-scale decentralized applications such as mobile keyboards, IoT analytics, and healthcare systems. By enabling multiple local communication rounds per selected cohort, our method significantly reduces communication overhead—a major bottleneck in practical deployments—without compromising model quality. As a result, *Cohort Squeeze* may make federated learning more accessible to organizations and regions with limited network infrastructure or compute capabilities.

The proposed method has the potential to enhance sustainability in distributed machine learning by reducing the number of global communication rounds, which often involve costly transmissions over wide-area networks. Additionally, our support for arbitrary client sampling strategies—including clustering and stratified approaches—offers flexibility to adapt to heterogeneous and evolving client populations.

However, as with any advancement in FL, there are considerations regarding fairness, privacy, and potential misuse. While our method is compatible with existing privacy-preserving mechanisms such as secure aggregation, we do not directly address differential privacy or robustness to adversarial behavior. Deploying such systems in sensitive domains (e.g., healthcare or finance) should be accompanied by safeguards ensuring that communication-efficient learning does not exacerbate performance disparities across client groups.

Overall, this work aims to advance the practicality and inclusiveness of federated learning, while encouraging the community to consider responsible deployment in real-world systems.

B Related Work

B.1 Cross-Device Federated Learning

This paper delves into the realm of Federated Learning (FL), focusing on the cross-device variant, which presents unique and significant challenges. In FL, two predominant settings are recognized: cross-silo and cross-device scenarios, as detailed in Table 1 of Kairouz et al., 2019. The primary distinction lies in the nature of the clients: cross-silo FL typically involves various organizations holding substantial data, whereas cross-device FL engages a vast array of mobile or IoT devices. In cross-device FL, the complexity is heightened by the inability to maintain a persistent hidden state for each client, unlike in cross-silo environments. This factor renders certain approaches impractical, particularly those reliant on stateful clients participating consistently across all rounds. Given the sheer volume of clients in cross-device FL, formulating and analyzing outcomes in an expectation form is more appropriate, but more complex than in finite-sum scenarios.

The pioneering and perhaps most renowned algorithm in cross-device FL is FedAvg (McMahan et al., 2017) and implemented in applications like Google's mobile keyboard (Hard et al., 2018; Yang et al., 2018; Ramaswamy et al., 2019). However, it is noteworthy that popular accelerated training algorithms such as Scaffold (Karimireddy et al., 2020b) and ProxSkip (Mishchenko et al., 2022b) are not aligned with our focus due to their reliance on memorizing the hidden state for each client, which is applicable for cross-device FL. Our research pivots on a novel variant within the cross-device framework. Once the cohort are selected for each global communication round, these cohorts engage in what we term as 'local communications' multiple times. The crux of our study is to investigate whether increasing the number of local communication rounds can effectively reduce the total communication cost to converge to a targeted accuracy.

B.2 Stochastic Proximal Point Method

Our exploration in this paper centers on the Stochastic Proximal Point Method (SPPM), a method extensively studied for its convergence properties. Initially termed as the incremental proximal point method by Bertsekas (2011), it was shown to converge nonasymptotically under the assumption of Lipschitz continuity for each f_i. Following this, Ryu & Boyd (2016) examined the convergence rates of SPPM, noting its resilience to inaccuracies in learning rate settings, contrasting with the behavior of Stochastic Gradient Descent (SGD). Further developments in SPPM's application were seen in the works of Patrascu & Necoara (2018), who analyzed its effectiveness in constrained optimization, incorporating random projections. Asi & Duchi (2019) expanded the scope of SPPM by studying a generalized method, AProx, providing insights into its stability and convergence rates under convex conditions. The research by Asi et al. (2020) and Chadha et al. (2022) further extended these findings, focusing on minibatching and convergence under interpolation in the AProx framework.

In the realm of federated learning, particularly concerning non-convex optimization, SPPM is also known as FedProx, as discussed in works like those of Li et al. (2020a) and Yuan & Li (2022). However, it is noted that in non-convex scenarios, the performance of FedProx/SPPM in terms of convergence rates does not surpass that of SGD. Beyond federated learning, the versatility of SPPM is evident in its application to matrix and tensor completion such as in the work of Bumin & Huang (2021). Moreover, SPPM has been adapted for efficient implementation in a variety of optimization problems, as shown by Shtoff (2022). While non-convex SPPM analysis presents significant challenges, with a full understanding of its convex counterpart still unfolding, recent studies such as the one by Khaled & Jin (2023) have reported enhanced convergence by leveraging second-order similarity. Diverging from this approach, our contribution is the development of an efficient minibatch SPPM method SPPM-AS that shows improved results without depending on such assumptions. Significantly, we also provide the first empirical evidence that increasing local communication rounds in finding the proximal point can lead to a reduction in total communication costs.

B.3 Local Solvers

Table 2: Local optimizers for solving the proximal subproblem.

Setting	1st order	2nd order
Strongly-Convex	Conjugate Gradients (CG) Accelerated GD Local GD Scaffnew	BFGS AICN LocalNewton
Nonconvex	Mime-Adam FedAdam-AdaGrad FedSpeed	Apollo OASIS

In the exploration of local solvers for the SPPM-AS algorithm, the focus is on evaluating the performance impact of various inexact proximal solvers within federated learning settings, spanning both strongly convex and non-convex objectives. Here's a simple summary of the algorithms discussed:

- FedAdagrad-AdaGrad (Wang et al., 2021b): Adapts AdaGrad for both client and server sides within federated learning, introducing local and global corrections to address optimizer state handling and solution bias.
- BFGS (Broyden, 1967; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970): A quasi-Newton method that approximates the inverse Hessian matrix to improve optimization efficiency, particularly effective in strongly convex settings but with limitations in distributed implementations.
- AICN (Hanzely et al., 2022): Offers a global $O(1/k^2)$ convergence rate under a semi-strong self-concordance assumption, streamlining Newton's method without the need for line searches.
- LocalNewton (Bischoff et al., 2023): Enhances local optimization steps with second-order information and global line search, showing efficacy in heterogeneous data scenarios despite a lack of extensive theoretical grounding.
- Fed-LAMB (Karimi et al., 2022): Extends the LAMB optimizer to federated settings, incorporating layer-wise and dimension-wise adaptivity to accelerate deep neural network training.
- FedSpeed (Sun et al., 2023): Aims to overcome non-vanishing biases and client-drift in federated learning through prox-correction and gradient perturbation steps, demonstrating effectiveness in image classification tasks.
- Mime-Adam (Karimireddy et al., 2020a): Mitigates client drift in federated learning by integrating global optimizer states and an SVRG-style correction term, enhancing the adaptability of Adam to distributed settings.
- OASIS (Jahani et al., 2021): Utilizes local curvature information for gradient scaling, providing an adaptive, hyperparameter-light approach that excels in handling ill-conditioned problems.

Table 3: Theoretical summary

Hyperparameter	Control	Rate (T)	Neighborhood
γ	$\uparrow$	$\downarrow$	$\uparrow$
$\mathcal{S}$	$\tau_{\mathcal{S}} \uparrow^{(1)}$	$\downarrow$	$\downarrow$
	Stratified sampling optimal clustering instead of BS or NICE sampling	$\downarrow$	Lemma 2.5

$^{(1)}$ We define $\tau_{\mathcal{S}} := \mathbb{E}_{S\sim\mathcal{S}}\left[|S|\right]$.

- Apollo (Ma, 2020): A quasi-Newton method that dynamically incorporates curvature information, showing improved efficiency and performance over first-order methods in deep learning applications.

Each algorithm contributes uniquely to the landscape of local solvers in federated learning, ranging from enhanced adaptivity and efficiency to addressing specific challenges such as bias, drift, and computational overhead.

B.4 Relation to Hierarchical Federated Learning.

Hierarchical federated learning (HFL) introduces one or more intermediate aggregation layers between clients and the central server (e.g., base stations or edge servers) in order to reduce wide-area communication and exploit local connectivity (Abad et al., 2020; Hamdi et al., 2021; Yang, 2021). In contrast, our core analysis in Secs. 2-3.5 assumes the standard cross-device topology with a single logical server and stateless clients. Cohort Squeeze reuses a sampled cohort for multiple rounds of server-mediated communication, without maintaining persistent group models at intermediate nodes. Sec. 3.6 demonstrates that the same SPPM-AS update can be run at hubs in an HFL stack, so our framework is complementary to, rather than a replacement for, existing hierarchical architectures.

C Theoretical Overview and Recommendations

C.1 Parameter Control

We have explored the effects of changing the hyperparameters of SPPM-AS on its theoretical properties, as summarized in Table 3. This summary shows that as the learning rate increases, the number of iterations required to achieve a target accuracy decreases, though this comes with an increase in neighborhood size. Focusing on sampling strategies, for SPPM-NICE employing NICE sampling, an increase in the sampling size $\tau_{\mathcal{S}}$ results in fewer iterations (T) and a smaller neighborhood. Furthermore, given that stratified sampling outperforms both block sampling and NICE sampling, we recommend adopting stratified sampling, as advised by Lemma 2.5.

C.2 Comparison of Sampling Strategies

Full Sampling (FS). Let $S = [n]$ with probability 1. Then SPPM-AS applied to Equation (10) becomes PPM (Moreau, 1965; Martinet, 1970) for minimizing f. Moreover, in this case, we have $p_i = 1$ for all $i \in [n]$ and Equation (6) takes on the form

$$\mu_{\mathrm{AS}} = \mu_{\mathrm{FS}} := \frac{1}{n}\sum_{i=1}^{n} \mu_i, \quad \sigma^2_{\star,\mathrm{AS}} = \sigma^2_{\star,\mathrm{FS}} := 0.$$

Note that μ_{FS} is the strong convexity constant of f, and that the neighborhood size is zero, as we would expect.

Table 4: Arbitrary samplings comparison.

Setting/Requirement	μ_{AS}	$\sigma_{\star,\text{AS}}$
Full	$\frac{1}{n}\sum_{i=1}^{n}\mu_i$	0
Non-Uniform	$\min_i \frac{\mu_i}{np_i}$	$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{np_i}\left\|\nabla f_i\left(x_\star\right)\right\|^2$
Nice	$\min_{C\subseteq[n],\lvert C\rvert=\tau}\frac{1}{\tau}\sum_{i\in C}\mu_i$	$\sum_{C\subseteq[n],\lvert C\rvert=\tau}\frac{1}{\binom{n}{\tau}}\left\|\frac{1}{\tau}\sum_{i\in C}\nabla f_i\left(x_\star\right)\right\|^2$
Block	$\min_{j\in[b]}\frac{1}{nq_j}\sum_{i\in C_j}\mu_i$	$\sum_{j\in[b]}q_j\left\|\sum_{i\in C_j}\frac{1}{np_i}\nabla f_i\left(x_\star\right)\right\|^2$
Stratified	$\min_{\mathbf{i}_b\in\mathbf{C}_b}\sum_{j=1}^{b}\frac{\mu_{i_j}\lvert C_j\rvert}{n}$	$\sum_{\mathbf{i}_b\in\mathbf{C}_b}\left(\prod_{j=1}^{b}\frac{1}{\lvert C_j\rvert}\right)\left\|\sum_{j=1}^{b}\frac{\lvert C_j\rvert}{n}\nabla f_{i_j}\left(x_\star\right)\right\|^2$ Upper bound: $\frac{b}{n^2}\sum_{j=1}^{b}\lvert C_j\rvert^2\sigma_j^2$

Nonuniform Sampling (NS). Let $S = \{i\}$ with probability $p_i > 0$, where $\sum_i p_i = 1$. Then Equation (6) takes on the form

$$\mu_{\text{AS}} = \mu_{\text{NS}} := \min_i \frac{\mu_i}{np_i}, \quad \sigma^2_{\star,\text{AS}} = \sigma^2_{\star,\text{NS}} := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{np_i}\left\|\nabla f_i\left(x_\star\right)\right\|^2.$$

If we take $p_i = \frac{\mu_i}{\sum_{j=1}^{n}\mu_j}$ for all $i \in [n]$, we shall refer to Algorithm 1 as SPPM with importance sampling (SPPM-IS). In this case,

$$\mu_{\text{NS}} = \mu_{\text{IS}} := \frac{1}{n}\sum_{i=1}^{n}\mu_i, \quad \sigma^2_{\star,\text{NS}} = \sigma^2_{\star,\text{IS}} := \frac{\sum_{i=1}^{n}\mu_i}{n}\sum_{i=1}^{n}\frac{\left\|\nabla f_i\left(x_\star\right)\right\|^2}{n\mu_i}.$$

This choice maximizes the value of μ_{NS} (and hence minimizes the first part of the convergence rate) over the choice of the probabilities.

Table 4 summarizes the parameters associated with various sampling strategies, serving as a concise overview of the methodologies discussed in the main text. This summary facilitates a quick comparison and reference.

C.3 Extreme Cases of Block Sampling and Stratified Sampling

Extreme cases of block sampling. We now consider two extreme cases:

- If $b = 1$, then SPPM-BS = SPPM-FS = PPM. Let's see, as a sanity check, whether we recover the right rate as well. We have $q_1 = 1, C_1 = [n], p_i = 1$ for all $i \in [n]$, and the expressions for μ_{AS} and $\sigma^2_{\star,\text{ BS}}$ simplify to

  $$\mu_{\text{BS}} = \mu_{\text{FS}} := \frac{1}{n}\sum_{i=1}^{n}\mu_i, \sigma^2_{\star,\text{BS}} = \sigma^2_{\star,\text{FS}} := 0.$$

  So, indeed, we recover the same rate as SPPM-FS.

- If $b = n$, then SPPM-BS = SPPM-NS. Let's see, as a sanity check, whether we recover the right rate as well. We have $C_i = \{i\}$ and $q_i = p_i$ for all $i \in [n]$, and the expressions for μ_{AS} and $\sigma^2_{\star,\text{BS}}$ simplify to

  $$\mu_{\text{BS}} = \mu_{\text{NS}} := \min_{i\in[n]}\frac{\mu_i}{np_i}, \quad \sigma^2_{\star,\text{BS}} = \sigma^2_{\star,\text{NS}} := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{np_i}\left\|\nabla f_i\left(x_\star\right)\right\|^2.$$

  So, indeed, we recover the same rate as SPPM-NS.

Extreme cases of stratified sampling. We now consider two extreme cases:

- If $b = 1$, then SPPM-SS = SPPM-US. Let's see, as a sanity check, whether we recover the right rate as well. We have $C_1 = [n], |C_1| = n, \left(\prod_{j=1}^{b} \frac{1}{|C_j|}\right) = \frac{1}{n}$ and hence
$$\mu_{\mathrm{SS}} = \mu_{\mathrm{US}} := \min_i \mu_i, \quad \sigma^2_{\star,\mathrm{SS}} = \sigma^2_{\star,\mathrm{US}} := \frac{1}{n}\sum_{i=1}^{n} \left\|\nabla f_i\left(x_\star\right)\right\|^2 .$$
So, indeed, we recover the same rate as SPPM-US.
- If $b = n$, then SPPM-SS = SPPM-FS. Let's see, as a sanity check, whether we recover the right rate as well. We have $C_i = \{i\}$ for all $i \in [n]$, $\left(\prod_{j=1}^{b} \frac{1}{|C_j|}\right) = 1$, and hence
$$\mu_{\mathrm{SS}} = \mu_{\mathrm{FS}} := \frac{1}{n}\sum_{i=1}^{n} \mu_i, \quad \sigma^2_{\star,\mathrm{SS}} = \sigma^2_{\star,\mathrm{FS}} := 0.$$
So, indeed, we recover the same rate as SPPM-FS.

C.4 Federated Averaging SPPM Baselines

In this section we propose two new algorithms based on Federated Averaging principle. Since to the best of our knowledge there are no federated averaging analyses within the same assumptions, we provide analysis of modified versions of SPPM-AS.

Averaging on $\mathrm{prox}_{\gamma f_i}$. We introduce FedProx-SPPM-AS (see Algorithm 2), which is inspired by the principles of FedProx (Li et al., 2020a). Unlike the traditional approach where a proximal operator is computed for the chosen cohort as a whole, in FedProx-SPPM-AS, we compute and then average the proximal operators calculated for each member within the cohort. However, this algorithm is not a simple case of SPPM-AS because it does not directly estimate the proximal operator at each step.

Algorithm 2 Proximal Averaging SPPM-AS (FedProx-SPPM-AS)

```
Input: starting point $x_{0,0} \in \mathbb{R}^d$, arbitrary sampling distribution $\mathcal{S}$, learning rate $\gamma > 0$, local communication rounds K.
for $t = 0, 1, 2, \cdots, T-1$ do
    Sample $S_t \sim \mathcal{S}$
    for $k = 0, 1, 2, \cdots K-1$ do
        $x_{k+1,t} = \sum_{i \in S_t} \frac{1}{|S_t|} \mathrm{prox}_{\gamma f_i}(x_{k,t})$
    end for
    $x_{0,t+1} \leftarrow x_{K,t}$
end for
Output: $x_{0,T}$
```

Algorithm 3 Federated Averaging SPPM-AS (FedAvg-SPPM-AS)

```
Input: starting point $x_{0,0} \in \mathbb{R}^d$, arbitrary sampling distribution $\mathcal{S}$, global learning rate $\gamma > 0$, local learning rate $\alpha > 0$, local communication rounds $K$
for $t = 0, 1, 2, \cdots, T-1$ do
    Sample $S_t \sim \mathcal{S}$
    $\forall i \in S_t \ \tilde{f}_{i,t}(x) \leftarrow f_i(x) + \frac{1}{2\gamma}\|x - x_t\|^2$
    for $k = 0, 1, 2, \cdots K-1$ do
        $x_{k+1,t} = \sum_{i \in S_t} \frac{1}{|S_t|} \mathrm{prox}_{\alpha \tilde{f}_{i,t}}(x_{k,t})$
    end for
    $x_{0,t+1} \leftarrow x_{K,t}$
end for
Output: $x_{0,T}$
```

Here, we employ a proof technique similar to that of Theorem 2.4 and obtain the following convergence.

Theorem C.1 (FedProx-SPPM-AS convergence)**.** *Let the number of local iterations $K = 1$, and assume that Assumption 2.1 (differentiability) and Assumption 2.2 (strong convexity) hold. Let $x_0 \in \mathbb{R}^d$ be an arbitrary starting point. Then, for any $t \geq 0$ and any $\gamma > 0$, the iterates of* FedProx-SPPM *(as described in Algorithm 2) satisfy:*

$$\mathbb{E}\left[\left\|x_t - x_\star\right\|^2\right] \leq A_{\mathcal{S}}^t \left\|x_0 - x_\star\right\|^2 + \frac{B_{\mathcal{S}}}{1 - A_{\mathcal{S}}},$$

where $A_{\mathcal{S}} := \mathbb{E}[[]\, S_t \sim \mathcal{S}] \frac{1}{|S_t|} \sum_{i \in S_t} \frac{1}{1+\gamma\mu_i}$ *and* $B_{\mathcal{S}} := \mathbb{E}[[]\, S_t \sim \mathcal{S}] \frac{1}{|S_t|} \sum_{i \in S_t} \frac{\gamma}{(1+\gamma\mu_i)\mu_i} \|\nabla f_i(x_\star))\|^2$.

Federated averaging for prox **approximation.** An alternative method involves estimating the proximal operator by averaging the proximal operators calculated for each worker's function. We call it *Federated Averaging Stochastic Proximal Point Method* (FedAvg-SPPM-AS, see Algorithm 3). (FedAvg-SPPM-AS, see Algorithm 3).

After selecting and fixing a sample of workers S_k, the main objective is to calculate the proximal operator. This can be accomplished by approximating the proximal calculation with the goal of minimizing $\tilde{f}_S(x) = f_S(x) + \frac{2}{\gamma}\left\|x - x_t\right\|^2$. It can be observed that this approach is equivalent to FedProx-SPPM-AS, as at each local step we calculate

$$\mathrm{prox}_{\alpha \tilde{f}_i}(x_{k,t}) := \arg\min_{z\in\mathbb{R}^d}\left[\tilde{f}_i(z) + \frac{2}{\alpha}\|z - x_{k,t}\|^2\right] = \arg\min_{z\in\mathbb{R}^d}\left[f_i(z) + \left(\frac{2}{\gamma} + \frac{2}{\alpha}\right)\|z - x_{k,t}\|^2\right].$$

D Training Details

D.1 Compute and Implementation Details

Our implementation builds on two open-source frameworks: Scafflix (Yi et al., 2025) for logistic regression experiments and FedLab (Zeng et al., 2023) for neural network experiments. All experiments were conducted on a single NVIDIA A100 GPU with 80GB of memory. Each configuration (e.g., different solvers, learning rates, or sampling strategies) was executed independently on this compute node.

Given the moderate size of the datasets and the communication-efficient nature of our method, the total computational cost is relatively low. We did not observe significant resource bottlenecks during training or evaluation.

All implementation details, including environment setup, command-line arguments, and scripts for reproducing the experiments, are provided in the supplementary code package. Instructions cover both convex and non-convex settings.

Further information on data usage, preprocessing, and non-IID partitioning is included in the following subsections.

All experiments can be reproduced using the provided scripts. For each figure in the paper we include a corresponding configuration file and a shell script (e.g., `run_fig1.sh`) that launches the exact training run with the reported hyperparameters and random seed. The top-level `README` in the code release documents the environment setup and commands for regenerating all plots.

D.2 System Assumptions and Implementation Model

Deployment model. We target cross-device FL with stateless clients and short within-round sessions (seconds to a few minutes). Unless stated, the topology is a single server (flat FL). In the hierarchical variant, a hub aggregates a cohort and relays to the server; communication accounting follows the communication-cost model at the end of Sec. 2.2.

Round structure and notation. At global round t, the server samples a cohort $C_t \subseteq [n]$ of size C, broadcasts x_t, and runs an inner loop of up to K *intra-cohort synchronizations* ("subrounds"). Let $C_{t,k}$ be the live set at subround $k \in \{0, \ldots, K-1\}$, with size $m_k := |C_{t,k}|$. The inner loop may terminate early at subround k and we then set $x_{t+1} \leftarrow x_{t,k}$; the number of executed subrounds is $K_{\text{eff}} \leq K$. The server-side update in each subround uses only updates received before a fixed timeout; late updates are ignored.

Within-round availability (quorum) and survival. We enforce a quorum $\tau \in (0,1]$ so that the inner loop proceeds only while $m_k \geq \tau C$. This yields an "elastic-K" mechanism that adapts K_{eff} to availability and prevents degenerate subrounds. In stress-tests we also consider an i.i.d. survival model in which each active client independently survives a subround with probability $s \in (0,1]$; thus $\mathbb{E}[m_k] \approx sC$ while the quorum guarantees $m_k \geq \tau C$. Our churn-robust recursion in App. G is stated for arbitrary shrinking sequences $C_{t,0} \supseteq \cdots \supseteq C_{t,K_{\text{eff}}-1}$, hence it covers both stochastic survival and worst-case drops subject to the quorum.

Communication-cost accounting. In flat FL each subround counts as one communication, so total cost for reaching accuracy ε is $C_{\text{flat}}(\varepsilon, K) = K\,T(\varepsilon)$. In hierarchical FL the per-subround and per-round costs differ; we use $C_{\text{hier}}(\varepsilon, K) = (c_1 K + c_2)T(\varepsilon)$ where c_1 is client↔hub and c_2 is hub↔server cost. Our bounds and plots report these metrics.

Churn-robust guarantees (summary; Appendix G.12). Let $\underline{\mu}$ and $\overline{\sigma}^2$ be lower/upper envelopes of the arbitrary-sampling constants over all live sets $C_{t,k}$. App. G shows that one round of SPPM-AS satisfies

$$\mathbb{E}\|x_{t+1} - x^\star\|^2 \le (1+\gamma\underline{\mu})^{-2K_{\text{eff}}}\|x_t - x^\star\|^2 + \frac{\gamma}{\underline{\mu}}\left(\overline{\sigma}^2 + \kappa B\rho^{K_{\text{eff}}} + \kappa' \sum_{k=0}^{K_{\text{eff}}-1} \frac{\sigma^2_{\text{sub}}}{m_k}\right).$$

With the quorum $m_k \ge \tau C$ the churn term is bounded by $\sum_k \frac{\sigma^2_{\text{sub}}}{m_k} \le \frac{K_{\text{eff}}}{\tau C}\sigma^2_{\text{sub}}$, so the iteration-complexity and total-cost bounds remain valid and the communication-cost objective still has a finite minimizer $K^\star$.

Learning-rate and prox-parameter choices under churn. For a live set of size m at a subround, the expected-smoothness stepsize for one GD synchronization on $f_{C_{t,k}}$ is

$$\eta^\star(m) = \frac{1}{\widetilde{L}_m}, \qquad \widetilde{L}_m = \frac{(n-m)L_{\max} + n(m-1)L_{\text{avg}}}{n-1}.$$

In practice we use a robust linear activity scaling $\eta(s) = \eta_0 \cdot \frac{m}{C} = \eta_0\, s$ with η_0 tuned at full participation ($m = C$). For LocalGD-K-Reuse we equalize the per-round move by $\eta(s, K) = \eta_0\, s/K$. For SPPM-AS we analogously scale the prox parameter as $\gamma(s) = \gamma_0\, s$; the main theorem holds for any $\gamma > 0$, and the iteration-complexity bound guides the range.

Stragglers, dropouts, and weighting. Each subround aggregates only received updates at the timeout; clients that time out are dropped for that subround and remain eligible in future rounds. Unless stated, aggregation is uniform over the live set $C_{t,k}$. The survival analysis in the appendix reports the measured K_{eff} for different (s, τ) and confirms that K_{eff} saturates under churn, which matches the theory that depends on K_{eff} rather than K.

Sampling and fairness across rounds. Across global rounds we use the arbitrary-sampling distributions defined in Sec. 2.4 (NICE, block, stratified). Clients are stateless, hence eligibility is independent across rounds; reuse is confined to the K_{eff} subrounds within a single global round.

Privacy and security. The protocol does not introduce additional information beyond standard FL. No raw data leave devices; only model updates are exchanged. Differential privacy or secure aggregation are orthogonal and can be layered on top.

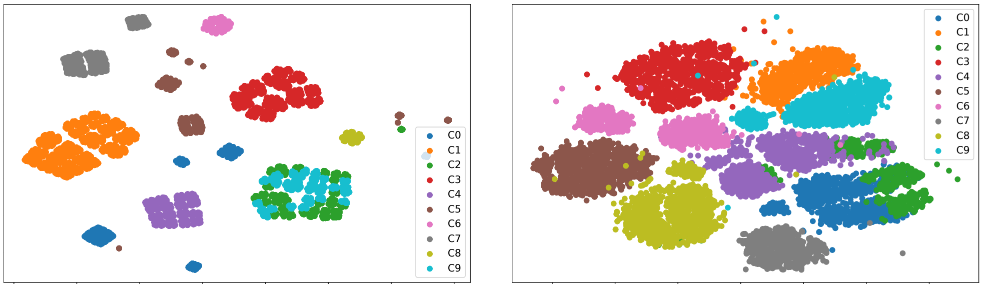
D.3 Non-IID Data Generation

In our study, we validate performance and compare the benefits of SPPM-AS over SPPM using well-known datasets such as `mushrooms`, `a6a`, `w6a`, and `ijcnn1.bz2` from LibSVM (Chang & Lin, 2011). To ensure relevance to our research focus, we adopt a feature-wise non-IID setting, characterized by variation in feature distribution across clients. This variation is introduced by clustering the features using the K-means algorithm, with the number of clusters set to 10 and the number of clients per cluster fixed at 10 for simplicity. We visualize the clustered data using t-SNE in Figure 5, where we observe that the data are divided into 10 distinct clusters with significantly spaced cluster centers.

D.4 Sampling

To simulate random sampling among clients within these 10 clusters, where each cluster comprises 10 clients, we consider two contrasting scenarios:

- *Case I* - SPPM-BS*:* Assuming clients within the same cluster share similar features and data distributions, sampling all clients from one cluster (i.e., $C = 10$ clients) results in a homogeneous sample.



(a) mushrooms, 10 clusters

(b) a6a, 10 clusters

Figure 5: t-SNE visualization of cluster-features across data samples on clients.

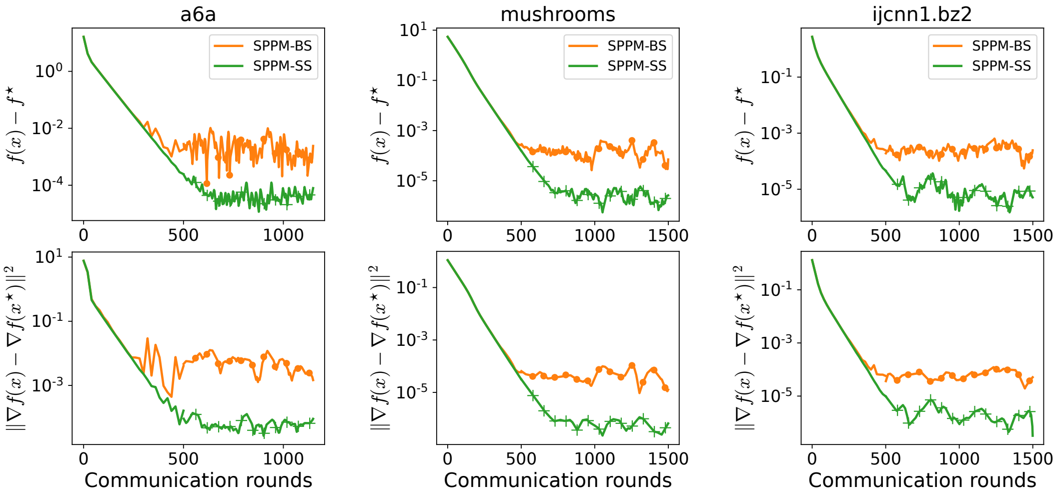


Figure 6: Comparison with SPPM-SS and SPPM-BS samplings.

- *Case II* - SPPM-SS*:* Conversely, by traversing all 10 clusters and randomly sampling one client from each, we obtain a group of 10 clients representing maximum heterogeneity.

We hypothesize that any random sampling from the 100 clients will yield performance metrics lying between these two scenarios. In Figure 6, we examine the impact of sampling clients with varying degrees of heterogeneity using a fixed learning rate of 0.1. Our findings indicate that heterogeneous sampling results in a significantly smaller convergence neighborhood $\sigma_\star^2$. This outcome is attributed to the broader global information captured through heterogeneous sampling, in contrast to homogeneous sampling, which increases the data volume without contributing additional global insights. As these two sampling strategies represent the extremes of arbitrary sampling, any random selection will fall between them in terms of performance. Given their equal cost and the superior performance of the SPPM-SS strategy in heterogeneous FL environments, we designate SPPM-SS as our default sampling approach.

D.5 SPPM-AS Algorithm Adaptation for Federated Learning

In the main text, Algorithm 1 outlines the general form of SPPM-AS. For the convenience of implementation in FL contexts and to facilitate a better understanding, we introduce a tailored version of the SPPM-AS algorithm specific to FL, designated as Algorithm 4. Notably, as block sampling is adopted as our default method, this adaptation of the algorithm specifically addresses the nuances of the block sampling approach. We also conducted arbitrary sampling on synthetic datasets and neural networks to demonstrate the algorithm's versatility.

Algorithm 4 SPPM-AS Adaptation for Federated Learning

```
Input: initial model $x_0 \in \mathbb{R}^d$; cohort size $C \geq 1$; sampling rule $S \in \{\text{NICE}, \text{BS}, \text{SS}\}$; local solver $\mathcal{A}$; prox stepsize $\gamma > 0$; number of local communication rounds $K \geq 1$.
for $t = 0, 1, 2, \dots$ do
    Sample cohort $C_t$ according to $S$:
        if $S = \text{BS}$: server samples a block $q_i$ and then $C$ clients from this block to form $C_t$
        else if $S = \text{SS}$: server samples $C$ blocks and one client from each selected block to form $C_t$
        else if $S = \text{NICE}$: server samples $C$ clients uniformly at random from $[n]$ to form $C_t$
    Server broadcasts the current global model $x_t$ to all clients $i \in C_t$
    Initialize cohort iterate $x_{t,0} \leftarrow x_t$
    for $k = 0, 1, \dots, K-1$ do
        Each client $i \in C_t$ computes a local update (e.g., gradient or step of $\mathcal{A}$) for the proximal subproblem $f_{C_t}(x) + \frac{1}{2\gamma}\|x - x_t\|^2$ at $x_{t,k}$ and sends it to the cohort aggregator
        Cohort aggregator applies one step of the solver $\mathcal{A}$ using the received updates and obtains a new shared iterate $x_{t,k+1}$
        Aggregator broadcasts $x_{t,k+1}$ back to all clients in $C_t$
    end for
    Server sets $x_{t+1} \leftarrow x_{t,K}$ (an inexact realization of $\text{prox}_{\gamma f_{C_t}}(x_t)$)
end for
```

D.6 Implementation of Local Communication Rounds

In our cross-device implementation, all communication is still mediated by the central server. At outer iteration t the server samples a cohort C_t and broadcasts the current model x_t to the clients in C_t. Each "local communication round" $k = 1, \dots, K$ then performs the following steps: (i) each client $i \in C_t$ runs τ local SGD steps on its objective f_i starting from the current server model; (ii) clients upload their updated parameters to the server; and (iii) the server aggregates these updates (by weighted averaging) to produce the next iterate $x_{t,k+1}$, which is again broadcast to C_t. No peer-to-peer communication between clients is used; the K rounds are simply K repeated server–cohort synchronizations on the same cohort, which together form an inexact evaluation of the cohort prox in SPPM-AS.

In a hierarchical deployment, client–hub communication can be significantly cheaper than hub–server communication. To capture this asymmetry we refine the cost model to

$$\text{Cost} = T\big(C_{\text{global}} + K C_{\text{local}}\big),$$

where C_{local} is the cost of one client–hub round and C_{global} is the cost of one hub–server round. The TK metric used in Secs. 3.3-3.5 corresponds to the special case $C_{\text{local}} = C_{\text{global}} = 1$, i.e., a single-hop cross-device topology with a central server.

E Additional Experiments on Logistic Regression

E.1 Communication Cost on Various Datasets to a Target Accuracy

In Figure 1, we presented the total communication cost relative to the number of rounds required to achieve the target accuracy for the selected cohort. In this section, we provide more details on how is this figure was obtained and present additional results for various datasets.

E.2 Convergence Speed and $\sigma^2_{\star,\text{SS}}$ Trade-Off

Unlike SGD-type methods such as MB-GD and MB-LocalGD, in which the largest allowed learning rate is $1/A$, where A is a constant proportion to the smoothness of the function we want to optimize (Gower et al., 2019). For larger learning rate, SGD-type method may not converge and exploding. However, for stochastic proximal point methods, they have a very descent benefit of allowing arbitrary learning rate. In this section, we verify whether our proposed method can allow

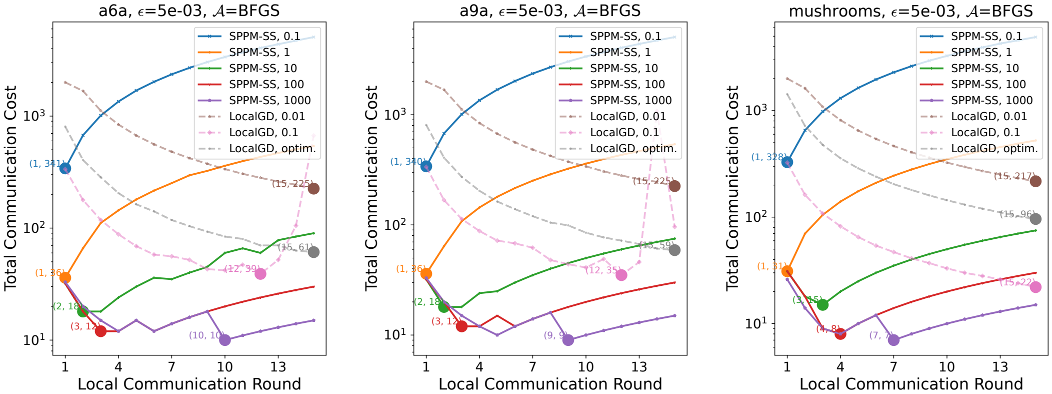


Figure 7: Total communication cost with respect to the local communication round. For LocalGD, K represents the local communication round K for finding the prox of the current model. For LocalGD, we slightly abuse the x-axis, which represents the total number of local iterations, no local communication is required. We calculate the total communication cost to reach a fixed global accuracy ϵ such that $\|x_t - x_\star\|^2 < \epsilon$. LocalGD, optim represents using the theoretical optimal stepsize of LocalGD with minibatch sampling.

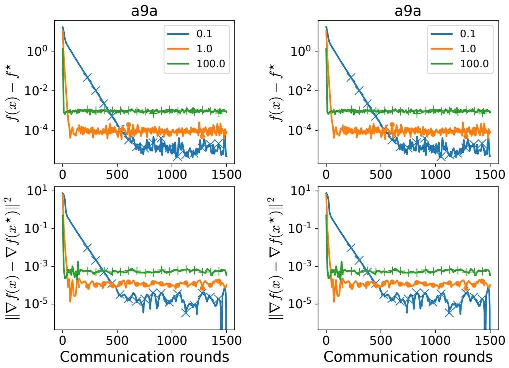


Figure 8: $K = 4$.

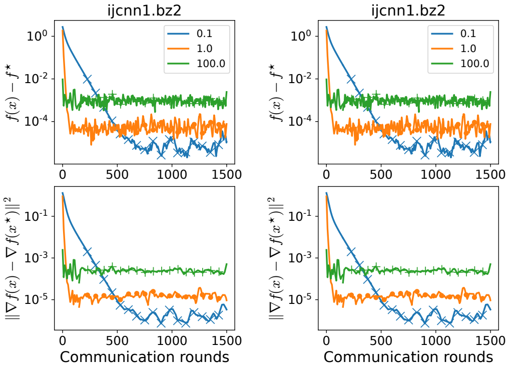


Figure 9: $K = 16$.

arbitrary learning rate and whether we can find something interesting. We considered different learning rate scale from 1e-5 to 1e+5. We randomly selected three learning rates [0.1, 1, 100] for visual representation with the results presented in Figure 8 and Figure 9. We found that a larger learning rate leads to a faster convergence rate but results in a much larger neighborhood, $\sigma^2_{\star,\mathrm{SS}}/\mu^2_{\mathrm{SS}}$. This can be considered a trade-off between convergence speed and neighborhood size, $\sigma^2_{\star,\mathrm{SS}}$. By default, we consider setting the learning rate to 1.0 which has a good balance between the convergence speed and the neighborhood size.

In this section, we extend our analysis by providing additional results across a broader range of datasets and varying learning rates. Specifically, Figure 8 illustrates the outcomes using 4 local communication rounds ($K = 4$), while Figure 9 details the results for 16 local communication rounds ($K = 16$). Previously, in Figure 1, we explored the advantages of larger K values. Here, our focus shifts to determining if similar trends are observable across different K values. Through comprehensive evaluations on various datasets and multiple K settings, we have confirmed that lower learning rates in SPPM-AS result in slower convergence speeds; however, they also lead to a smaller final convergence neighborhood.

E.3 Additional Experiments on Hierarchical Federated Learning

In Figure 2d of the main text, we detail the total communication cost for hierarchical Federated Learning (FL) utilizing parameters $c_1 = 0.1$ and $c_2 = 1$ on the `a6a` dataset. Our findings reveal that SPPM-AS achieves a significant reduction in communication costs, amounting to 94.87%, compared with the conventional FL setting where $c_1 = 1$ and $c_2 = 1$, which shows a 74.36% reduction. In this section, we extend our analysis with comprehensive evaluations on additional datasets, namely `ijcnn1.bz2`, `a9a`, and `mushrooms`. Beyond considering $c_1 = 0.1$, we further explore the impact of reducing the local communication cost from each client to the corresponding hub to $c_1 = 0.05$. The results, presented in Figure 10 and the continued Figure 11, reinforce our observation: hierarchical FL consistently leads to further reductions in communication costs. A lower c_1 parameter correlates with even greater savings in communication overhead. These results not only align with our expectations but also underscore the efficacy of our proposed SPPM-AS in cross-device FL settings.

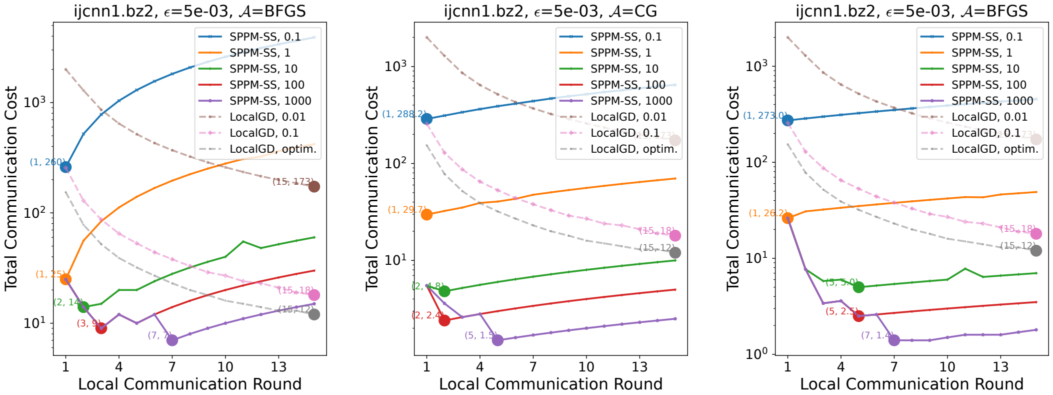
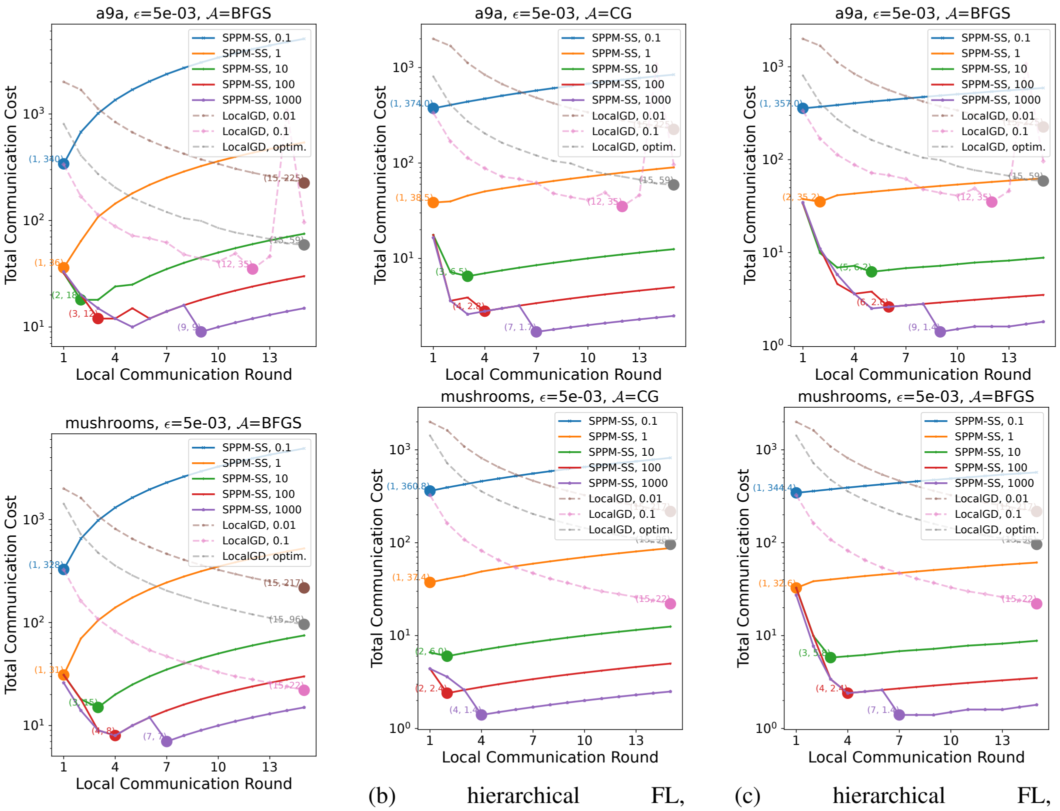


Figure 10: The total communication cost is analyzed with respect to the number of local communication rounds. For LocalGD, K represents the local communication round used for finding the prox of the current model. In the case of LocalGD, we slightly abuse the x-axis to represent the total number of local iterations, as no local communication is required. We calculate the total communication cost needed to reach a fixed global accuracy ϵ, such that $\|x_t - x_\star\|^2 < \epsilon$. LocalGD, optim denotes the use of the theoretically optimal stepsize for LocalGD with minibatch sampling. Comparisons are made between different prox solvers (CG and BFGS).



(a) standard FL, $c_1 = 1, c_2 = 0$

(b) hierarchical FL, $c_1 = 0.1, c_2 = 1$

(c) hierarchical FL, $c_1 = 0.05, c_2 = 1$

Figure 11: Total communication cost with respect to the local communication round.

Table 5: Architecture of the CNN model for FEMNIST symbol recognition.

Layer	Output Shape	# of Trainable Parameters	Activation	Hyperparameters
Input	(28, 28, 1)	0		
Conv2d	(24, 24, 32)	832	ReLU	kernel size = 5; strides = (1, 1)
Conv2d	(10, 10, 64)	51,264	ReLU	kernel size = 5; strides = (1, 1)
MaxPool2d	(5, 5, 64)	0		pool size = (2, 2)
Flatten	6400	0		
Dense	128	819,328	ReLU	
Dense	62	7,998	softmax	

F Additional Neural Network Experiments

F.1 Experiment Details

For our neural network experiments, we used the FEMNIST dataset (Caldas et al., 2018). Each client was created by uniformly selecting from user from original dataset, inherently introducing heterogeneity among clients. We tracked and reported key evaluation metrics—training and testing loss and accuracy—after every 5 global communication rounds. The test dataset was prepared by dividing each user's data into a 9:1 ratio, following the partitioning approach of the FedLab framework (Zeng et al., 2023). For the SPPM-AS algorithm, we selected Adam as the optimizer for the proximal operator. The learning rate was determined through a grid search across the following range: $[0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5]$. The model architecture comprises a convolutional neural network (CNN) with the following layers: Conv2d(1, 32, 5), ReLU, Conv2d(32, 64, 5), MaxPool2d(2, 2), a fully connected (FC) layer with 128 units, ReLU, and another FC layer with 128 units, as specified in Table 5. Dropout, learning rate scheduling, gradient clipping, etc., were not used to improve the interpretability of results.

We explore various values of targeted training accuracy, as illustrated in Figure 12. This analysis helps us understand the impact of different accuracy thresholds on the model's performance. For instance, we observe that as the target accuracy changes, SPPM-NICE consistently outperforms LocalGD in terms of total communication cost. As the target accuracy increases, the performance gap between these two algorithms also widens. Additionally, we perform ablation studies on different values of c_1, as shown in Figure 13, to assess their effects on the learning process. Here, we note that with $c_2 = 0.2$, SPPM-NICE performs similarly to LocalGD, suggesting that an increase in c_2 value could narrow the performance gap between SPPM-NICE and LocalGD.

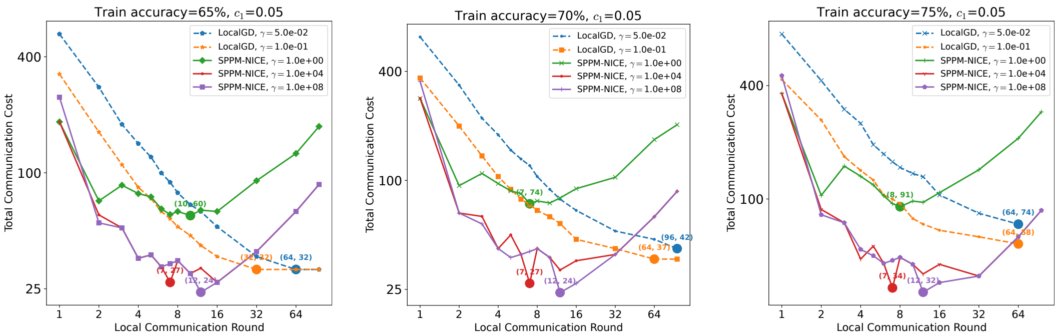


Figure 12: Varying targeted training accuracy level for SPPM-AS.

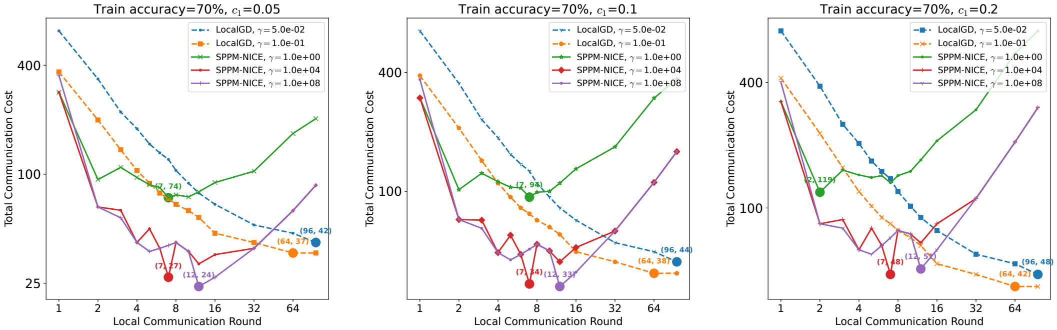


Figure 13: Varying c_1 cost.

F.2 Convergence Analysis Compared with Baselines

Further, we compare SPPM-AS, SPPM, and LocalGD in Figure 15, placing a particular emphasis on evaluating the total computational complexity. This measure gains importance in scenarios where communication rounds are of secondary concern, thereby shifting the focus to the assessment of computational resource expenditure.

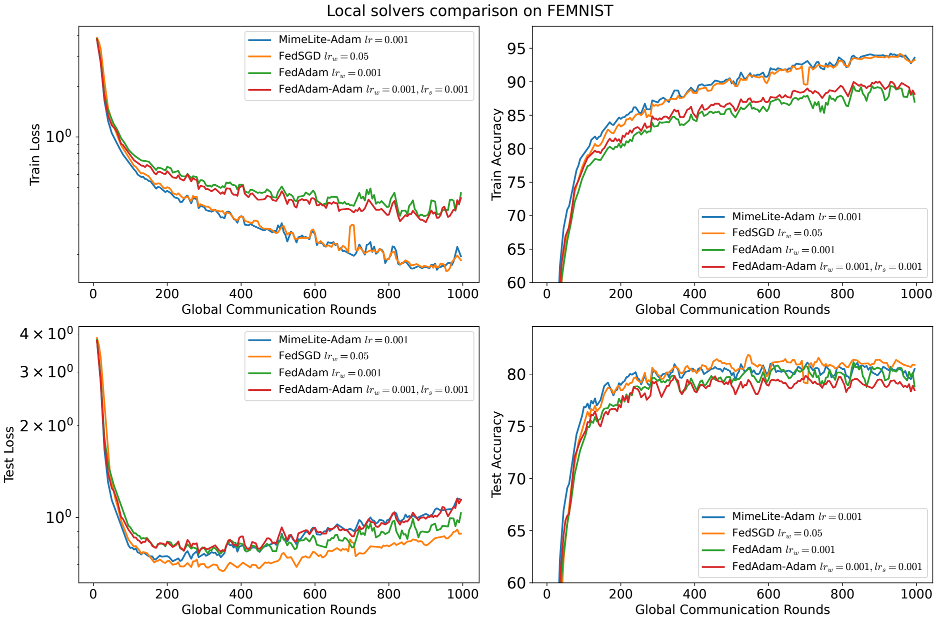


Figure 14: Different local solvers for prox baselines for training a CNN model over 100 workers using data from the FEMNIST dataset. The number of local communication rounds is fixed at 3 and the number of worker optimizer steps is fixed at 3. Nice sampling with a minibatch size of 10 is used. γ is fixed at 1.0.

F.3 Prox Solvers Baselines

We compare baselines from B.3 for training a CNN model over 100 workers using data from the FEMNIST dataset, as shown in Figure 14. The number of local communication rounds and worker optimizer steps is consistent among various solvers for the purpose of fair comparison. All local solvers optimize the local objective, which is prox on the selected cohort. The solvers compared are: LocalGD referred as FedSGD (McMahan et al., 2017) - the Federated Averaging algorithm with SGD as the worker optimizer, FedAdam - the Federated Averaging algorithm with Adam as the worker optimizer, FedAdam-Adam based on the FedOpt framework (Reddi et al., 2020), and finally

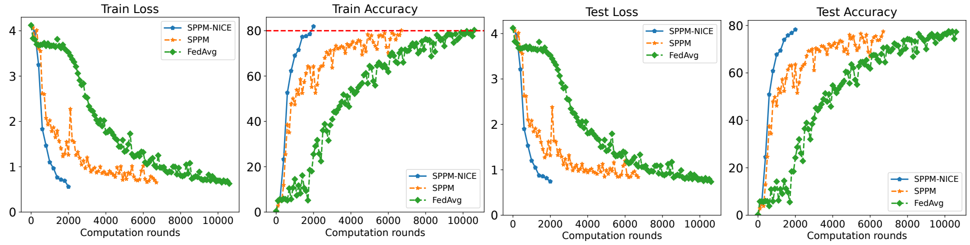


Figure 15: Accuracy compared with baselines.

MimeLite-Adam, which is based on the Mime (Karimireddy et al., 2020a) framework and the Adam optimizer. The hyperparameter search included a double-level sweep of the optimizer learning rates: $[0.00001, 0.0001, 0.001, 0.01, 0.1]$, followed by $[0.25, 0.5, 1.0, 2.5, 5] * lr_{\text{best}}$. One can see that all methods perform similarly, with MimeLite-Adam and FedSGD converging better on the test data.

G Missing Proof and Additional Theoretical Analysis

G.1 Facts Used in the Proof

Fact G.1 (Differentiation of integral with a parameter (theorem 2.27 from Folland (1984))). *Suppose that $f : X \times [a, b] \to \mathbb{C}(-\infty < a < b < \infty)$ and that $f(\cdot, t) : X \to \mathbb{C}$ is integrable for each $t \in [a, b]$. Let $F(t) = \int_X f(x, t) d\mu(x)$.*

1. *Suppose that there exists $g \in L^1(\mu)$ such that $|f(x, t)| \leq g(x)$ for all x, t. If $\lim_{t\to t_0} f(x, t) = f(x, t_0)$ for every x, then $\lim_{t\to t_0} F(t) = F(t_0)$; in particular, if $f(x, \cdot)$ is continuous for each x, then F is continuous.*
2. *Suppose that $\partial f/\partial t$ exists and there is a $g \in L^1(\mu)$ such that $|(\partial f/\partial t)(x, t)| \leq g(x)$ for all x, t. Then F is differentiable and $F'(x) = \int (\partial f/\partial t)(x, t) d\mu(x)$.*

Fact G.2 (Tower Property). *For any random variables X and Y, we have*

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]\,.$$

Fact G.3 (Every point is a fixed point (Khaled & Jin, 2023)). *Let $\varphi : \mathbb{R}^d \to \mathbb{R}$ be a convex differentiable function. Then*

$$\operatorname{prox}_{\gamma\varphi}(x + \gamma\nabla\varphi(x)) = x, \qquad \forall \gamma > 0, \quad \forall x \in \mathbb{R}^d.$$

In particular, if $x_\star$ is a minimizer of φ, then $\operatorname{prox}_{\gamma\varphi}(x_\star) = x_\star$.

Proof. Evaluating the proximity operator is equivalent to

$$\operatorname{prox}_{\gamma\varphi}(y) = \arg\min_{x\in\mathbb{R}^d} \left(\varphi(x) + \frac{1}{2\gamma}\|x - y\|^2\right).$$

This is a strongly convex minimization problem for any $\gamma > 0$, hence the (necessarily unique) minimizer $x = \operatorname{prox}_{\gamma\varphi}(y)$ of this problem satisfies the first-order optimality condition

$$\nabla\varphi(x) + \frac{1}{\gamma}(x - y) = 0.$$

Solving for y, we observe that this holds for $y = x + \gamma\nabla\phi(x)$. Therefore, $x = \operatorname{prox}_{\gamma\varphi}(x + \gamma\nabla\varphi(x))$. □

Fact G.4 (Contractivity of the prox (Mishchenko et al., 2022a)). *If φ is differentiable and μ-strongly convex, then for all $\gamma > 0$ and for any $x, y \in \mathbb{R}^d$ we have*

$$\left\|\operatorname{prox}_{\gamma\varphi}(x) - \operatorname{prox}_{\gamma\varphi}(y)\right\|^2 \leq \frac{1}{(1 + \gamma\mu)^2}\|x - y\|^2\,.$$

Fact G.5 (Recurrence (Khaled & Jin, 2023, Lemma 1))**.** *Assume that a sequence $\{s_t\}_{t\geq 0}$ of positive real numbers for all $t \geq 0$ satisfies*

$$s_{t+1} \leq a s_t + b,$$

where $0 < a < 1$ and $b \geq 0$. Then the sequence for all $t \geq 0$ satisfies

$$s_t \leq a^t s_0 + b \min\left\{t, \frac{1}{1-a}\right\}.$$

Proof. Unrolling the recurrence, we get

$$s_t \leq a s_{t-1} + b \leq a(a s_{t-2} + b) + b \leq \cdots \leq a^t s_0 + b \sum_{i=0}^{t-1} a^i.$$

We can now bound the sum $\sum_{i=0}^{t-1} a^i$ in two different ways. First, since $a < 1$, we get the estimate

$$\sum_{i=0}^{t-1} a^i \leq \sum_{i=0}^{t-1} 1 = t.$$

Second, we sum a geometic series

$$\sum_{i=0}^{t-1} a^i \leq \sum_{i=0}^{\inf} a^i = \frac{1}{1-a}.$$

Note that either of these bounds can be better. So, we apply the best of these bounds. Substituing the above two bounds gived the target inequality. □

G.2 Simplified Proof of SPPM

We provide a simplified proof of SPPM (Khaled & Jin, 2023) in this section. Using the fact that $x_\star = \mathrm{prox}_{\gamma f_{\xi_t}}(x_\star + \gamma \nabla f_{\xi_t}(x_\star))$ (see Fact G.3) and then applying contraction of the prox (Fact G.4), we get

$$\begin{aligned}
\|x_{t+1} - x_\star\|^2 &= \left\|\mathrm{prox}_{\gamma f_{\xi_t}} - x_\star\right\|^2 \\
&\overset{(Fact\ G.3)}{=} \left\|\mathrm{prox}_{\gamma f_{\xi_t}}(x_t) - \mathrm{prox}_{\gamma f_{\xi_t}}(x_\star + \gamma \nabla f_{\xi_t}(x_\star))\right\|^2 \\
&\overset{(Fact\ G.4)}{\leq} \frac{1}{(1+\gamma\mu)^2} \|x_t - (x_\star + \gamma \nabla f_{\xi_t}(x_\star))\|^2 \\
&= \frac{1}{(1+\gamma\mu)^2} \left(\|x_t - x_\star\|^2 - 2\gamma \langle \nabla f_{\xi_t}(x_\star), x_t - x_\star\rangle + \gamma^2 \|\nabla f_{\xi_t}(x_\star)\|^2\right).
\end{aligned}$$

Taking expectation on both sides, conditioned on x_t, we get

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1} - x_\star\|^2 | x_t\right] &\leq \frac{1}{(1+\gamma\mu)^2} \left(\|x_t - x_\star\|^2 - 2\gamma \langle \mathbb{E}[\nabla f_{\xi_t}(x_\star)], x_t - x_\star\rangle + \gamma^2 \mathbb{E}\left[\|\nabla f_{\xi_t}(x_\star)\|^2\right]\right) \\
&= \frac{1}{(1+\gamma\mu)^2} \left(\|x_t - x_\star\|^2 + \gamma^2 \sigma_\star^2\right),
\end{aligned}$$

where we used the fact that $\mathbb{E}[\nabla f_{\xi_t}(x_\star)] = \nabla f(x_\star) = 0$ and $\sigma_\star^2 := \mathbb{E}\left[\|\nabla f_{\xi_t}(x_\star)\|^2\right]$. Taking expectation again and applying the tower property (Fact G.2), we get

$$\mathbb{E}\left[\|x_{t+1} - x_\star\|^2\right] \leq \frac{1}{(1+\gamma\mu)^2} \left(\|x_t - x_\star\|^2 + \gamma^2 \sigma_\star^2\right).$$

It only remains to solve the above recursion. Luckily, that is exactly what Fact G.5 does. In particular, we use it with $s_t = \mathbb{E}\left[\|x_t - x_\star\|^2\right]$, $a = \frac{1}{(1+\gamma\mu)^2}$ and $b = \frac{\gamma^2\sigma_\star^2}{(1+\gamma\mu)^2}$ to get

$$\begin{aligned}\mathbb{E}\left[\|x_t - x_\star\|^2\right] &\overset{(Fact\ G.5)}{\leq} \left(\frac{1}{1+\gamma\mu}\right)^{2t} \|x_0 - x_\star\|^2 + \frac{\gamma^2\sigma_\star^2}{(1+\gamma\mu)^2} \min\left\{t, \frac{(1+\gamma\mu)^2}{(1+\gamma\mu)^2 - 1}\right\} \\ &\leq \left(\frac{1}{1+\gamma\mu}\right)^{2t} \|x_0 - x_\star\|^2 + \frac{\gamma^2\sigma_\star^2}{(1+\gamma\mu)^2 - 1} \\ &\leq \left(\frac{1}{1+\gamma\mu}\right)^{2t} \|x_0 - x_\star\|^2 + \frac{\gamma\sigma_\star^2}{\gamma\mu^2 + 2\mu}.\end{aligned}$$

G.3 Missing Proof of Theorem 2.4

We first prove the following useful lemma.

Lemma G.6. *Let $\phi_\xi : \mathbb{R}^d \to \mathbb{R}$ be differentiable functions for almost all $\xi \sim \mathcal{D}$, with ϕ_ξ being μ_ξ-strongly convex for almost all $\xi \sim \mathcal{D}$. Further, let w_ξ be positive scalars. Then the function $\phi := \mathbb{E}[[]\ \xi \sim \mathcal{D}] w_\xi \phi_\xi$ is μ-strongly convex with $\mu = \mathbb{E}[[]\ \xi \sim \mathcal{D}] w_\xi \mu_\xi$.*

Proof. By assumption,

$$\phi_\xi(y) + \langle \nabla\phi_\xi(y), x - y\rangle + \frac{\mu_\xi}{2}\|x - y\|^2 \leq \phi_\xi(x), \quad \text{for almost all } \xi \in \mathcal{D}, \forall x, y \in \mathbb{R}^d.$$

This means that

$$\mathbb{E}[[]\ \xi \sim \mathcal{D}] w_\xi \left(\phi_\xi(y) + \langle \nabla\phi_\xi(y), x - y\rangle + \frac{\mu_\xi}{2}\|x - y\|^2\right) \leq \mathbb{E}[[]\ \xi \sim \mathcal{D}] w_\xi \phi_\xi(x), \quad \forall x, y \in \mathbb{R}^d,$$

which is equivalent to

$$\phi(y) + \langle \nabla\phi(y), x - y\rangle + \frac{\mathbb{E}[[]\ \xi \sim \mathcal{D}] w_\xi \mu_\xi}{2}\|x - y\|^2 \leq \phi(x), \quad \forall x, y \in \mathbb{R}^d,$$

So, ϕ is μ-strongly convex. □

Now, we are ready to prove our main Theorem 2.4.

Proof. Let C be any (necessarily nonempty) subset of $[n]$ such that $p_C > 0$. Recall that in view of Equation (9) we have

$$f_C(x) = \mathbb{E}[[]\ \xi \sim \mathcal{D}] \frac{I\left(\xi \in C\right)}{p_\xi} f_\xi(x)$$

i.e., f_C is a conic combination of the functions $\{f_\xi : \xi \in C\}$ with weights $w_\xi = \frac{I(\xi \in C)}{p_\xi}$. Since each f_ξ is μ_ξ-strongly convex, Lemma G.6 says that f_C is μ_C-strongly convex with

$$\mu_C := \mathbb{E}[[]\ \xi \sim \mathcal{D}] \frac{I\left(\xi \in C\right)\mu_\xi}{p_\xi}.$$

So, every such f_C is μ-strongly convex with

$$\mu = \mu_{\text{AS}} := \min_{C \subseteq [n], p_C > 0} \mathbb{E}[[]\ \xi \sim \mathcal{D}] \frac{I\left(\xi \in C\right)\mu_\xi}{p_\xi}.$$

Further, the quantity $\sigma_\star^2$ from (2.3) is equal to

$$\sigma_\star^2 := \mathrm{E}_{\xi \sim \mathcal{D}}\left[\|\nabla f_\xi\left(x_\star\right)\|^2\right] \overset{Eqn.\ (11)}{=} \sum_{C \subseteq [n], p_C > 0} p_C \|\nabla f_C\left(x_\star\right)\|^2 := \sigma_{\star,\text{AS}}^2.$$

Incorporating Appendix G.2 into the above equation, we prove the theorem. □

G.4 Theory for Expectation Formulation

We will formally define our optimization objective, focusing on minimization in expectation form. We consider

$$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E}[[]\, \xi \sim \mathcal{D}] f_\xi(x), \tag{7}$$

where $f_\xi : \mathbb{R}^d \to \mathbb{R}$, $\xi \sim \mathcal{D}$ is a random variable following distribution $\mathcal{D}$.

Assumption G.7. *Function $f_\xi : \mathbb{R}^d \to \mathbb{R}$ is differentiable for almost all samples $\xi \sim \mathcal{D}$.*

This implies that f is differentiable. We will implicitly assume that the order of differentiation and expectation can be swapped [1], which means that

$$\nabla f(x) \overset{Eqn.\ (1)}{=} \nabla \mathbb{E}[[]\, \xi \sim \mathcal{D}] f_\xi(x) = \mathbb{E}[[]\, \xi \sim \mathcal{D}] \nabla f_\xi(x).$$

Assumption G.8. *Function $f_\xi : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex for almost all samples $\xi \sim \mathcal{D}$, where $\mu > 0$. That is*

$$f_\xi(y) + \langle \nabla f_\xi, x - y \rangle + \frac{\mu}{2} \|x - y\|^2 \leq f_\xi(x),$$

for all $x, y \in \mathbb{R}^d$.

This implies that f is μ-strongly convex, and hence f has a unique minimizer, which we denote by $x_\star$. We know that $\nabla f(x_\star) = 0$. Notably, we do *not* assume f to be L-smooth.

Let $\mathcal{S}$ be a probability distribution over all *finite* subsets of $\mathbb{N}$. Given a random set $S \sim \mathcal{S}$, we define

$$p_i := \mathrm{Prob}(i \in S), \quad i \in \mathbb{N}.$$

We will restrict our attention to proper and nonvacuous random sets.

Assumption G.9. *S is proper (i.e., $p_i > 0$ for all $i \in \mathbb{N}$) and nonvacuous (i.e., $\mathrm{Prob}(S = \emptyset) = 0$).*

Let C be the selected cohort. Given $\emptyset \neq C \subset \mathbb{N}$ and $i \in \mathbb{N}$, we define

$$v_i(C) := \begin{cases} \frac{1}{p_i} & i \in C \\ 0 & i \notin C, \end{cases} \tag{8}$$

and

$$f_C(x) := \mathbb{E}[[]\, \xi \sim \mathcal{D}] v_\xi(C) f_\xi(x) \overset{Eqn.\ (8)}{=} \mathbb{E}[[]\, \xi \sim \mathcal{D}] \frac{I\,(\xi \in C)}{p_\xi} f_\xi(x). \tag{9}$$

Note that $v_i(S)$ is a random variable and f_S is a random function. By construction, $\mathrm{E}_{S \sim \mathcal{S}}\left[v_i(S)\right] = 1$ for all $i \in \mathbb{N}$, and hence

$$\begin{aligned} \mathbb{E}[[]\, S \sim \mathcal{S}] f_S(x) &= \mathbb{E}[[]\, S \sim \mathcal{S}] \mathbb{E}[[]\, \xi \sim \mathcal{D}] v_\xi(C) \nabla f_\xi(x) \\ &= \mathbb{E}[[]\, \xi \sim \mathcal{D}] \mathbb{E}[[]\, S \sim \mathcal{S}] v_\xi(S) \nabla f_\xi(x) = \mathbb{E}[[]\, \xi \sim \mathcal{D}] f_\xi(x) = f(x). \end{aligned}$$

Therefore, the optimization problem in Equation (1) is equivalent to the stochastic optimization problem

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) := \mathrm{E}_{S \sim \mathcal{S}}\left[f_S(x)\right] \right\}. \tag{10}$$

Further, if for each $C \subset \mathbb{N}$ we let $p_C := \mathrm{Prob}(S = C)$, f can be written in the equivalent form

$$f(x) = \mathbb{E}[[]\, S \sim \mathcal{S}] f_S(x) = \sum_{C \subset \mathbb{N}} p_C f_C(x) = \sum_{C \subset \mathbb{N}, p_C > 0} p_C f_C(x). \tag{11}$$

[1] This assumption satisfies the conditions required for the theorem about differentiating an integral with a parameter (Fact G.1).

Theorem G.10 (Main Theorem)**.** *Let Assumption 2.1 (diferentiability) and Assumption 2.2 (strong convexity) hold. Let S be a random set satisfying Assumption 2.3, and define*

$$\mu_{\mathrm{AS}} := \min_{C\subset\mathbb{N},p_C>0} \mathbb{E}[[]\,\xi\sim\mathcal{D}]\frac{I\left(\xi\in C\right)\mu_\xi}{p_\xi},$$
$$\sigma^2_{\star,\mathrm{AS}} := \sum_{C\subset\mathbb{N},p_C>0} p_C\left\|\nabla f_C\left(x_\star\right)\right\|^2. \tag{12}$$

Let $x_0\in\mathbb{R}^d$ be an arbitrary starting point. Then for any $t\geq 0$ and any $\gamma>0$, the iterates of SPPM-AS *(Algorithm 1) satisfy*

$$\mathrm{E}\left[\|x_t-x_\star\|^2\right]\leq\left(\frac{1}{1+\gamma\mu_{\mathrm{AS}}}\right)^{2t}\|x_0-x_\star\|^2+\frac{\gamma\sigma^2_{\star,\mathrm{AS}}}{\gamma\mu^2_{\mathrm{AS}}+2\mu_{\mathrm{AS}}}.$$

G.5 Missing Proof of Iteration Complexity of SPPM-AS

We have seen above that accuracy arbitrarily close to (but not reaching) $\sigma^2_{\star,\mathrm{AS}}/\mu^2_{\mathrm{AS}}$ can be achieved via a single step of the method, provided the stepsize γ is large enough. Assume now that we aim for ϵ accuracy where $\epsilon\leq\sigma^2_{\star,\mathrm{AS}}/\mu^2_{\mathrm{AS}}$. Using the inequality $1-k\leq\exp(-k)$ which holds for all $k>0$, we get

$$\left(\frac{1}{1+\gamma\mu_{\mathrm{AS}}}\right)^{2t}=\left(1-\frac{\gamma\mu}{1+\gamma\mu_{\mathrm{AS}}}\right)^{2t}\leq\exp\left(-\frac{2\gamma\mu_{\mathrm{AS}}t}{1+\gamma\mu_{\mathrm{AS}}}\right)$$

Therefore, provided that

$$t\geq\frac{1+\gamma\mu_{\mathrm{AS}}}{2\gamma\mu_{\mathrm{AS}}}\log\left(\frac{2\left\|x_0-x_\star\right\|^2}{\varepsilon}\right),$$

we get $\left(\frac{1}{1+\gamma\mu_{\mathrm{AS}}}\right)^{2t}\|x_0-x_\star\|^2\leq\frac{\varepsilon}{2}$. Furthermore, as long as $\gamma\leq\frac{2\varepsilon\mu_{\mathrm{AS}}}{2\sigma^2_{\star,\mathrm{AS}}-\varepsilon\mu^2_{\mathrm{AS}}}$ (this is true provided that the more restrictive but also more elegant-looking condition $\gamma\leq\varepsilon\mu_{\mathrm{AS}}/\sigma^2_{\star,\mathrm{AS}}$ holds), we get $\frac{\gamma\sigma^2_{\star,\mathrm{AS}}}{\gamma\mu^2_{\mathrm{AS}}+2\mu_{\mathrm{AS}}}\leq\frac{\varepsilon}{2}$. Putting these observations together, we conclude that with the stepsize $\gamma=\varepsilon\mu_{\mathrm{AS}}/\sigma^2_{\star,\mathrm{AS}}$, we get $\mathrm{E}\left[\left\|x_t-x_\star\right\|^2\right]\leq\varepsilon$ provided that

$$t\geq\frac{1+\gamma\mu_{\mathrm{AS}}}{2\gamma\mu_{\mathrm{AS}}}\log\frac{2\left\|x_0-x_\star\right\|^2}{\varepsilon}=\left(\frac{\sigma^2_{\star,\mathrm{AS}}}{2\varepsilon\mu^2_{\mathrm{AS}}}+\frac{1}{2}\right)\log\left(\frac{2\left\|x_0-x_\star\right\|^2}{\varepsilon}\right).$$

G.6 $\sigma^2_{\star,\mathrm{NICE}}(\tau)$ and $\mu_{\mathrm{NICE}}(\tau)$ are Monotonous Functions of τ

Lemma G.11. *For all $0\leq\tau\leq n-1$:*

1. $\mu_{\mathrm{NICE}}(\tau+1)\geq\mu_{\mathrm{NICE}}(\tau)$,
2. $\sigma^2_{\star,\mathrm{NICE}}(\tau)=\frac{\frac{n}{\tau}-1}{n-1}\sigma^2_{\star,\mathrm{NICE}}(1)\leq\frac{1}{\tau}\sigma^2_{\star,\mathrm{NICE}}(1)$.

Proof. 1. Pick any $1\leq\tau<n$, and consider a set C for which the minimum is attained in

$$\mu_{\mathrm{NICE}}(\tau+1)=\min_{C\subseteq[n],|C|=\tau+1}\frac{1}{\tau+1}\sum_{i\in C}\mu_i.$$

Let $j=\arg\max_{i\in C}\mu_i$. That is, $\mu_j\geq\mu_i$ for all $i\in C$. Let C_j be the set obtained from C by removing the element j. Then $|C_j|=\tau$ and

$$\mu_j=\max_{i\in C}\mu_i\geq\max_{i\in C_j}\mu_i\geq\frac{1}{\tau}\sum_{i\in C_j}\mu_i.$$

By adding $\sum_{i\in C_j}\mu_i$ to the above inequality, we obtain

$$\mu_j + \sum_{i\in C_j}\mu_i \geq \frac{1}{\tau}\sum_{i\in C_j}\mu_i + \sum_{i\in C_j}\mu_i.$$

Observe that the left-hand side is equal to $\sum_{i\in C}\mu_i$, and the right-hand side is equal to $\frac{\tau+1}{\tau}\sum_{i\in C_j}\mu_i$. If we divide both sides by $\tau+1$, we obtain

$$\frac{1}{\tau+1}\sum_{i\in C}\mu_i \geq \frac{1}{\tau}\sum_{i\in C_j}\mu_i.$$

Since the left-hand side is equal to $\mu_{\mathrm{NICE}}(\tau+1)$, and the right hand side is an upper bound on $\mu_{\mathrm{NICE}}(\tau)$, we conclude that $\mu_{\mathrm{NICE}}(\tau+1) \geq \mu_{\mathrm{NICE}}(\tau)$.

2. In view of equation 9 we have

$$f_C(x) = \sum_{i\in C}\frac{1}{np_i}f_i(x). \tag{13}$$

$$\sigma^2_{\star,\mathrm{AS}} \quad = \quad \mathrm{E}_{S\sim\mathcal{S}}\left[\left\|\sum_{i\in S}\frac{1}{np_i}\nabla f_i(x_\star)\right\|^2\right] = \mathrm{E}_{S\sim\mathcal{S}}\left[\left\|\sum_{i\in S}\frac{1}{\tau}\nabla f_i(x_\star)\right\|^2\right] \tag{14}$$

Let χ_i be the random variable defined by

$$\chi_j = \begin{cases} 1 & j\in S \\ 0 & j\notin S. \end{cases} \tag{15}$$

It is easy to show that

$$\mathbb{E}[\chi_j] = \mathrm{Prob}(j\in S) = \frac{\tau}{n}. \tag{16}$$

Let fix the cohort S. Let χ_{ij} be the random variable defined by

$$\chi_{ij} = \begin{cases} 1 & i\in S \text{ and } j\in S \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Note that

$$\chi_{ij} = \chi_i\chi_j. \tag{18}$$

Further, it is easy to show that

$$\mathbb{E}[\chi_{ij}] = \mathrm{Prob}(i\in S, j\in S) = \frac{\tau(\tau-1)}{n(n-1)}. \tag{19}$$

Denote $a_i := \nabla f_i(x_\star)$.

$$
\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{\tau}\sum_{i\in S} a_i\right\|^2\right] &= \frac{1}{\tau^2}\mathbb{E}\left[\left\|\sum_{i\in S} a_i\right\|^2\right] \\
&= \frac{1}{\tau^2}\mathbb{E}\left[\left\|\sum_{i=1}^{n} \chi_i a_i\right\|^2\right] \\
&= \frac{1}{\tau^2}\mathbb{E}\left[\sum_{i=1}^{n} \|\chi_i a_i\|^2 + \sum_{i\neq j} \langle \chi_i a_i, \chi_j a_j\rangle\right] \\
&= \frac{1}{\tau^2}\mathbb{E}\left[\sum_{i=1}^{n} \|\chi_i a_i\|^2 + \sum_{i\neq j} \chi_{ij} \langle a_i, a_j\rangle\right] \\
&= \frac{1}{\tau^2}\sum_{i=1}^{n} \mathbb{E}[\chi_i] \|a_i\|^2 + \sum_{i\neq j} \mathbb{E}[\chi_{ij}] \langle a_i, a_j\rangle \\
&= \frac{1}{\tau^2}\left(\frac{\tau}{n}\sum_{i=1}^{n} \|a_i\|^2 + \frac{\tau(\tau-1)}{n(n-1)}\sum_{i\neq j} \langle a_i, a_j\rangle\right) \\
&= \frac{1}{\tau n}\sum_{i=1}^{n} \|a_i\|^2 + \frac{\tau-1}{\tau n(n-1)}\sum_{i\neq j} \langle a_i, a_j\rangle \\
&= \frac{1}{\tau n}\sum_{i=1}^{n} \|a_i\|^2 + \frac{\tau-1}{\tau n(n-1)}\left(\left\|\sum_{i=1}^{n} a_j\right\|^2 - \sum_{i=1}^{n} \|a_i\|^2\right) \\
&= \frac{n-\tau}{\tau(n-1)}\frac{1}{n}\sum_{i=1}^{n} \|a_i\|^2 + \frac{n(\tau-1)}{\tau(n-1)}\left\|\frac{1}{n}\sum_{i=1}^{n} a_i\right\|^2 \\
&= \frac{n-\tau}{\tau(n-1)}\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x_\star)\|^2 + \frac{n(\tau-1)}{\tau(n-1)}\left\|\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_\star)\right\|^2 \\
&= \frac{n-\tau}{\tau(n-1)}\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x_\star)\|^2 \\
&\leq \frac{1}{\tau}\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x_\star)\|^2
\end{aligned}
$$

□

G.7 MISSING PROOF OF LEMMA 2.5

For ease of notation, let $a_i = \nabla f_i\left(x_\star\right)$ and $\hat{z}_j = \left|C_j\right| a_{\xi_j}$, and recall that

$$
\sigma_{\star,\mathrm{SS}}^2 = \mathrm{E}_{\xi_1,\ldots,\xi_b}\left[\left\|\frac{1}{n}\sum_{j=1}^{b} \hat{z}_j\right\|^2\right]. \tag{20}
$$

where $\xi_j \in C_j$ is chosen uniformly at random. Further, for each $j \in [b]$, let $z_j = \sum_{i \in C_j} a_i$. Observe that $\sum_{j=1}^{b} z_j = \sum_{j=1}^{b} \sum_{i \in C_j} a_i = \sum_{i=1}^{n} a_i = \nabla f(x_\star) = 0$. Therefore,

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{j=1}^{b} \hat{z}_j \right\|^2 &= \frac{1}{n^2} \left\| \sum_{j=1}^{b} \hat{z}_j - \sum_{j=1}^{b} z_j \right\|^2 \\
&= \frac{b^2}{n^2} \left\| \frac{1}{b} \sum_{j=1}^{b} (\hat{z}_j - z_j) \right\|^2 \\
&\leq \frac{b^2}{n^2} \frac{1}{b} \sum_{j=1}^{b} \| \hat{z}_j - z_j \|^2 \\
&= \frac{b}{n^2} \sum_{j=1}^{b} \| \hat{z}_j - z_j \|^2 , \qquad (21)
\end{aligned}$$

where the inequality follows from convexity of the function $u \mapsto \|u\|^2$. Next,

$$\| \hat{z}_j - z_j \|^2 = \left\| |C_j| \, a_{\xi_j} - \sum_{i \in C_j} a_i \right\|^2 = |C_j|^2 \left\| a_{\xi_j} - \frac{1}{|C_j|} \sum_{i \in C_j} a_i \right\|^2 \leq |C_j|^2 \, \sigma_j^2. \qquad (22)$$

By combining Equation (20), Equation (21) and Equation (22), we get

$$\begin{aligned}
\sigma^2_{\star,\mathrm{SS}} &\overset{Eqn.\ (20)}{=} \mathrm{E}_{\xi_1,\ldots,\xi_b} \left[\left\| \frac{1}{n} \sum_{j=1}^{b} \hat{z}_j \right\|^2 \right] \\
&\overset{Eqn.\ (21)}{\leq} \mathrm{E}_{\xi_1,\ldots,\xi_b} \left[\frac{b}{n^2} \sum_{j=1}^{b} \| \hat{z}_j - z_j \|^2 \right] \\
&\overset{Eqn.\ (22)}{\leq} \mathrm{E}_{\xi_1,\ldots,\xi_b} \left[\frac{b}{n^2} \sum_{j=1}^{b} |C_j|^2 \, \sigma_j^2 \right] \\
&= \frac{b}{n^2} \sum_{j=1}^{b} |C_j|^2 \, \sigma_j^2 .
\end{aligned}$$

The last expression can be further bounded as follows:

$$\frac{b}{n^2} \sum_{j=1}^{b} |C_j|^2 \sigma_j^2 \leq \frac{b}{n^2} \left(\sum_{j=1}^{b} |C_j|^2 \right) \max_j \sigma_j^2 \leq \frac{b}{n^2} \left(\sum_{j=1}^{b} |C_j| \right)^2 \max_j \sigma_j^2 = b \max_j \sigma_j^2 ,$$

where the second inequality follows from the relation $\|u\|_2 \leq \|u\|_1$ between the L_2 and L_1 norms, and the last identity follows from the fact that $\sum_{j=1}^{b} |C_j| = n$.

G.8 Stratified Sampling Against Block Sampling and Nice Sampling

In this section, we present a theoretical comparison of block sampling and its counterparts, providing a theoretical justification for selecting block sampling as the default clustering method in future experiments. Additionally, we compare various sampling methods, all with the same sampling size, b: b-nice sampling, block sampling with b clusters, and block sampling, where all clusters are of uniform size b.

Assumption G.12. *For simplicity of comparison, we assume b clusters, each of the same size, b:*

$$|C_1| = |C_2| = \ldots = |C_b| = b.$$

It is crucial to acknowledge that, without specific assumptions, the comparison of different sampling methods may not provide meaningful insights. For instance, the scenario described in Lemma 2.5, characterized by complete inter-cluster homogeneity, demonstrates that block sampling achieves a variance term, denoted as $\sigma^2_{\star,\mathrm{SS}}$, which is lower than the variance terms associated with both block sampling and nice sampling. However, a subsequent example illustrates examples in which the variance term for block sampling surpasses those of block sampling and nice sampling.

Example G.13. *Without imposing any additional clustering assumptions, there exist examples for any arbitrary n, such that $\sigma^2_{\star,\mathrm{SS}} \geq \sigma^2_{\star,\mathrm{BS}}$ and $\sigma^2_{\star,\mathrm{SS}} \geq \sigma^2_{\star,\mathrm{NICE}}$.*

Proof. **Counterexample when SS is worse in neighborhood than BS**

Assume we have such clustering and $\nabla f_i(x_\star)$ such that the centroids of each cluster are equal to zero: $\forall i \in [b]$, $\frac{1}{|C_i|}\sum_{j\in C_i} \nabla f_j(x_\star) = 0$. For instance, this can be achieved in the following case: The dimension is $d = 2$, all clusters are of equal size m, then assign $\forall i \in [b]$, $\forall j \in C_i$, $\nabla f_j(x_\star) = \left(Re\left(\omega^{mj+i}\right), Im\left(\omega^{mj+i}\right)\right)$ where $\omega = \sqrt[n]{1} \in \mathbb{C}$. Let us calculate $\sigma^2_{\star,\mathrm{BS}}$:

$$\sigma^2_{\star,\mathrm{BS}} := \sum_{j=1}^{b} q_j \left\|\sum_{i\in C_j} \frac{1}{np_i}\nabla f_i(x_\star)\right\|^2 =$$

$$= \frac{1}{n^2}\sum_{j=1}^{b}\frac{|C_j|^2}{q_j}\left\|\frac{1}{|C_j|}\sum_{i\in C_j}\nabla f_i(x_\star)\right\|^2 = 0.$$

As a result:

$$\sigma^2_{\star,\mathrm{BS}} = 0 \leq \sigma^2_{\star,\mathrm{SS}}.$$

Counterexample when SS is worse in neighborhood than NICE

Here, we employ a similar proof technique as in the proof of Lemma 2.6. Let us choose such clustering $\mathcal{C}_{b,\mathrm{SS},\max} = \arg\max_{\mathcal{C}_b} \sigma^2_{\star,\mathrm{SS}}(\mathcal{C}_b)$. Denote $\mathbf{i}_b := (i_1, \cdots, i_b)$, $\mathbf{C}_b := C_1 \times \cdots \times C_b$, and $S_{\mathbf{i}_b} := \left\|\frac{1}{\tau}\sum_{i\in\mathbf{i}_b}\nabla f_i(x_\star)\right\|$.

$$\sigma^2_{\star,\mathrm{NICE}} = \frac{1}{C(n,\tau)}\sum_{C\subseteq[n],|C|=\tau}\left\|\frac{1}{\tau}\sum_{i\in C}\nabla f_i(x_\star)\right\|^2$$

$$= \frac{1}{C(n,b)}\sum_{\mathbf{i}_b\subseteq[n]} S_{\mathbf{i}_b}$$

$$\overset{1}{=} \frac{1}{\#_{\text{clusterizations}}}\sum_{\mathcal{C}_b}\frac{1}{b^b}\sum_{\mathbf{i}_b\in\mathbf{C}_b} S_{\mathbf{i}_b}$$

$$= \frac{1}{\#_{\text{clusterizations}}}\sum_{\mathcal{C}_b}\sigma^2_{\star,\mathrm{SS}}(\mathcal{C}_b)$$

$$\overset{2}{\leq} \sigma^2_{\star,\mathrm{SS}}(\mathcal{C}_{b,\mathrm{SS},\max}).$$

Equation 1 holds because, in every clusterization $\mathcal{C}_b$, there are $\frac{1}{b^b}$ possible sample combinations $\mathbf{i}_b$. Due to symmetry, one can conclude that each combination $S_{\mathbf{i}_b}$ is counted the same number of times. Equation 2 follows from the definition of $\mathcal{C}_{b,\mathrm{SS},\max}$.

For illustrative purposes, we can demonstrate this effect with a specific example. Let $n = 4$ and define $\forall i\ a_i = \nabla f_i(x^*) \in \mathbb{R}^2$. Let $a_1 = (0,1)^T$, $a_2 = (1,0)^T$, $a_3 = (0,-1)^T$, and $a_4 = (-1,0)^T$. Then fix clustering $\mathcal{C}_b = \{C_1 = \{a_1, a_3\}, C_2 = \{a_2, a_4\}\}$. Then:

$$\sigma^2_{\star,\mathrm{SS}} = \frac{1}{4}\sum_{\mathbf{i}_b\in\mathcal{C}_b}\left\|\frac{a_{i_1}+a_{i_2}}{2}\right\|^2$$

$$= \frac{1}{4}\sum_{\mathbf{i}_b\in\mathcal{C}_b}\left\|(\pm\frac{1}{2},\pm\frac{1}{2})\right\|^2$$

$$= \frac{1}{2}.$$

$$
\begin{aligned}
\sigma^2_{\star,\mathrm{NICE}} &= \frac{1}{C(4,2)} \sum_{i<j} \left\| \frac{a_i + a_j}{2} \right\|^2 \\
&= \frac{1}{6} \sum_{i<j} \left\| \frac{a_i + a_j}{2} \right\|^2 \\
&= \frac{1}{6} \left(\left[\left\| \frac{a_1 + a_3}{2} \right\|^2 + \left\| \frac{a_2 + a_4}{2} \right\|^2 \right] + 2 \times \left\| \frac{a_{i_1} + a_{i_2}}{2} \right\|^2 \right) \\
&= \frac{1}{6} \left(0 + 2 \times 2 \times \frac{1}{2} \right) \\
&= \frac{1}{3} \\
&= \frac{2}{3} \times \sigma^2_{\star,\mathrm{SS}} \\
&\le \sigma^2_{\star,\mathrm{SS}}
\end{aligned}
$$

□

To select the optimal clustering, we will choose the clustering that minimizes $\sigma^2_{\star,\mathrm{SS}}$.

Definition G.14 (Stratified sampling optimal clustering)**.** *Denote the clustering of workers into blocks as $\mathcal{C}_b := \{C_1, C_2, \ldots, C_b\}$, such that the disjoint union of all clusters $C_1 \cup C_2 \cup \ldots \cup C_b = [n]$. Define* block sampling Optimal Clustering *as the clustering configuration that minimizes $\sigma^2_{\star,\mathrm{SS}}$, formally given by:*

$$\mathcal{C}_{b,\mathrm{SS}} := \arg\min_{\mathcal{C}_b} \sigma^2_{\star,\mathrm{SS}}(\mathcal{C}_b).$$

Lemma G.15. *Given Assumption G.12, the following holds: $\sigma^2_{\star,\mathrm{SS}}\left(\mathcal{C}_{b,\mathrm{SS}}\right) \le \sigma^2_{\star,\mathrm{NICE}}$ for arbitrary b. Moreover, the variance within the convergence neighborhood of stratified sampling is less than or equal to that of nice sampling: $\frac{\gamma\sigma^2_{\star,\mathrm{SS}}}{\gamma\mu^2_{\mathrm{SS}}+2\mu_{\mathrm{SS}}}\left(\mathcal{C}_{b,\mathrm{SS}}\right) \le \frac{\gamma\sigma^2_{\star,\mathrm{NICE}}}{\gamma\mu^2_{\mathrm{NICE}}+2\mu_{\mathrm{NICE}}}$.*

Proof. 1. Denote $\mathbf{i}_b := (i_1, \cdots, i_b)$, $\mathbf{C}_b := C_1 \times \cdots \times C_b$, and $S_{\mathbf{i}_b} := \left\| \frac{1}{\tau} \sum_{i \in \mathbf{i}_b} \nabla f_i(x_\star) \right\|$.

$$
\begin{aligned}
\sigma^2_{\star,\mathrm{NICE}} &= \frac{1}{C(n,\tau)} \sum_{C \subseteq [n], |C| = \tau} \left\| \frac{1}{\tau} \sum_{i \in C} \nabla f_i(x_\star) \right\|^2 \\
&= \frac{1}{C(n,b)} \sum_{\mathbf{i}_b \subseteq [n]} S_{\mathbf{i}_b} \\
&\overset{1}{=} \frac{1}{\#\text{clusterizations}} \sum_{\mathcal{C}_b} \frac{1}{b^b} \sum_{\mathbf{i}_b \in \mathbf{C}_b} S_{\mathbf{i}_b} \\
&= \frac{1}{\#\text{clusterizations}} \sum_{\mathcal{C}_b} \sigma^2_{\star,\mathrm{SS}}(\mathcal{C}_b) \\
&\overset{2}{\ge} \sigma^2_{\star,\mathrm{SS}}(\mathcal{C}_{b,\mathrm{SS},\min})
\end{aligned}
$$

Equation 1 holds because, in every clusterization $\mathcal{C}_b$, there are $\frac{1}{b^b}$ possible sample combinations $\mathbf{i}_b$. Due to symmetry, one can conclude that each combination $S_{\mathbf{i}_b}$ is counted the same number of times. Equation 2 follows from the definition of $\mathcal{C}_{b,\mathrm{SS},\min}$ as the clustering that minimizes $\sigma^2_{\star,\mathrm{SS}}$, according to Definition G.14.

2. The neighborhood size for SPPM-AS is given by $\frac{\gamma\sigma^2_{\star,\mathrm{AS}}}{\gamma\mu^2_{\mathrm{AS}}+2\mu_{\mathrm{AS}}}$, denoted as U_{AS} for simplicity. Define:

$$\mu_{\mathrm{NICE}(b)} := \min_{\substack{C\subseteq[n]\\|C|=b}} \frac{1}{b}\sum_{i\in C}\mu_i,$$

$$\mu_{\mathrm{SS}} := \min_{\mathbf{i}_b\in\mathbf{C}_b}\sum_{j=1}^{b}\frac{\mu_{i_j}|C_j|}{n} \overset{\text{Asm. 10}}{=} \min_{\mathbf{i}_b\in\mathbf{C}_b}\sum_{j=1}^{b}\frac{\mu_{i_j}b}{b^2} = \min_{\mathbf{i}_b\in\mathbf{C}_b}\frac{1}{b}\sum_{j=1}^{b}\mu_{i_j}.$$

Using the definition of the set $\mathbf{C}_b := C_1\times C_2\times\cdots\times C_b$, we have $\mathbf{C}_b \subseteq \{C\subseteq[n] \mid |C| = b\}$. Applying this fact, we obtain:

$$\mu_{\mathrm{SS}} = \min_{\mathbf{i}_b\in\mathbf{C}_b}\frac{1}{b}\sum_{j\in\mathbf{i}_b}\mu_j \geq \mu_{\mathrm{NICE}(b)}.$$

Combining the above with $\sigma^2_{\star,\mathrm{SS}}\left(\mathcal{C}_{b,\mathrm{SS}}\right) \leq \sigma^2_{\star,\mathrm{NICE}}$, we obtain that $U_{\mathrm{SS}}\left(\mathcal{C}_{b,\mathrm{SS}}\right) \leq U_{\mathrm{NICE}}$, demonstrating the variance reduction of SS compared to NICE.

□

Example G.16. *Consider the number of clusters and the size of each cluster, with $b = 2$, under Assumption G.12. Then, $\sigma^2_{\star,\mathrm{SS}}\left(\mathcal{C}_{b,\mathrm{SS}}\right) \leq \sigma^2_{\star,\mathrm{BS}}$.*

Proof. Let $n = 4$, $b = 2$. Denote $\forall i\ \ a_i = \nabla f_i(x_*)$. Define $S^2 := \sum_{i<j}\left\|\frac{a_i+a_j}{2}\right\|^2$.

$$\begin{aligned}\sigma^2_{\star,\mathrm{SS}} &= \frac{1}{4}\left(S^2 - \left\|\frac{a_{C_1^1}+a_{C_1^2}}{2}\right\|^2 - \left\|\frac{a_{C_2^1}+a_{C_2^2}}{2}\right\|^2\right)\\ &= \frac{1}{4}\left(S^2 - 2\sigma^2_{\star,\mathrm{BS}}\right)\end{aligned}$$

$\mathcal{C}_{b,\mathrm{SS}}$ clustering minimizes $\sigma^2_{\star,\mathrm{SS}}$, thereby maximizing $\sigma^2_{\star,\mathrm{BS}}$. Thus,

$$\begin{aligned}\sigma^2_{\star,\mathrm{SS}} &= \frac{1}{4}\left(\left[\left\|\frac{a_{C_1^1}+a_{C_2^1}}{2}\right\|^2 + \left\|\frac{a_{C_1^2}+a_{C_2^2}}{2}\right\|^2\right] + \left[\left\|\frac{a_{C_1^1}+a_{C_2^2}}{2}\right\|^2 + \left\|\frac{a_{C_1^2}+a_{C_2^1}}{2}\right\|^2\right]\right)\\ &= \frac{1}{4}\left(2\sigma^2_{\star,\mathrm{BS}}\left((C_1^1,C_2^1),(C_1^2,C_2^2)\right) + 2\sigma^2_{\star,\mathrm{BS}}\left((C_1^1,C_2^2),(C_1^2,C_2^1)\right)\right)\\ &= \frac{1}{2}\left(\sigma^2_{\star,\mathrm{BS}}\left((C_1^1,C_2^1),(C_1^2,C_2^2)\right) + \sigma^2_{\star,\mathrm{BS}}\left((C_1^1,C_2^2),(C_1^2,C_2^1)\right)\right)\\ &\leq \sigma^2_{\star,\mathrm{BS}}.\end{aligned}$$

□

However, it is possible that this relationship might hold more generally. Empirical experiments for different configurations, such as $b = 3$, support this possibility. For example, with $n = 9$, $b = 3$, and $d = 10$, Python simulations where gradients ∇f_i are sampled from $\mathcal{N}(0,1)$ and $\mathcal{N}(e,1)$ across 1000 independent trials, show that $\sigma^2_{\star,\mathrm{SS}} \leq \sigma^2_{\star,\mathrm{BS}}$. Question of finding theoretical proof for arbitraty n remains open and has yet to be addressed in the existing literature.

G.9 Different Approaches of Federated Averaging

Proof of Theorem C.1:

Proof.

$$
\begin{aligned}
\|x_t - x_\star\|^2 &= \left\| \sum_{i \in S_t} \frac{1}{|S_t|} \operatorname{prox}_{\gamma f_i}(x_{t-1}) - \frac{1}{|S_t|} \sum_{i \in S_t} x_\star \right\|^2 \\
&\overset{(Fact\ G.3)}{=} \left\| \sum_{i \in S_t} \frac{1}{|S_t|} \left[\operatorname{prox}_{\gamma f_i}(x_{t-1}) - \operatorname{prox}_{\gamma f_i}(x_\star + \gamma \nabla f_i(x_\star)) \right] \right\|^2 \\
&\overset{Jensen}{\leq} \sum_{i \in S_t} \frac{1}{|S_t|} \left\| \left[\operatorname{prox}_{\gamma f_i}(x_{t-1}) - \operatorname{prox}_{\gamma f_i}(x_\star + \gamma \nabla f_i(x_\star)) \right] \right\|^2 \\
&\overset{(Fact\ G.4)}{\leq} \sum_{i \in S_t} \frac{1}{|S_t|} \frac{1}{(1+\gamma\mu_i)^2} \|x_{t-1} - (x_\star + \gamma \nabla f_i(x_\star))\|^2
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}[[]\, S_t \sim \mathcal{S}] \|x_t - x_\star\|^2 | x_{t-1} \\
&\leq \mathbb{E}[[]\, S_t \sim \mathcal{S}] \sum_{i \in S_t} \frac{1}{|S_t|} \frac{1}{(1+\gamma\mu_i)^2} \|(x_{t-1} - x_\star) - \gamma \nabla f_i(x_\star))\|^2 | x_{t-1} \\
&\overset{\text{Young}, \alpha_i > 0}{\leq} \mathbb{E}[[]\, S_t \sim \mathcal{S}] \sum_{i \in S_t} \frac{1}{|S_t|} \frac{1}{(1+\gamma\mu_i)^2} \left((1+\alpha_i) \|x_{t-1} - x_\star\|^2 + \left(1 + \alpha_i^{-1}\right) \|\gamma \nabla f_i(x_\star))\|^2 \right) | x_{t-1} \\
&\overset{\alpha_i = \gamma\mu_i}{=} \mathbb{E}[[]\, S_t \sim \mathcal{S}] \sum_{i \in S_t} \frac{1}{|S_t|} \frac{1}{(1+\gamma\mu_i)^2} \left((1+\gamma\mu_i) \|x_{t-1} - x_\star\|^2 + \left(1 + \frac{1}{\gamma\mu_i}\right) \|\gamma \nabla f_i(x_\star))\|^2 \right) | x_{t-1} \\
&= \mathbb{E}[[]\, S_t \sim \mathcal{S}] \sum_{i \in S_t} \frac{1}{|S_t|} \left(\frac{1}{1+\gamma\mu_i} \|x_{t-1} - x_\star\|^2 + \frac{\gamma}{(1+\gamma\mu_i)\mu_i} \|\nabla f_i(x_\star))\|^2 \right) | x_{t-1} \\
&= \mathbb{E}[[]\, S_t \sim \mathcal{S}] \frac{1}{|S_t|} \sum_{i \in S_t} \frac{1}{1+\gamma\mu_i} | x_{t-1} \|x_{t-1} - x_\star\|^2 + \mathbb{E}[[]\, S_t \sim \mathcal{S}] \frac{1}{|S_t|} \sum_{i \in S_t} \frac{\gamma}{(1+\gamma\mu_i)\mu_i} \|\nabla f_i(x_\star))\|^2 | x_{t-1}
\end{aligned}
$$

By applying tower property one can get the following:

$$
\begin{aligned}
&\mathbb{E}[[]\, S_t \sim \mathcal{S}] \|x_t - x_\star\|^2 \\
&= \mathbb{E}[[]\, S_t \sim \mathcal{S}] \frac{1}{|S_t|} \sum_{i \in S_t} \frac{1}{1+\gamma\mu_i} \|x_{t-1} - x_\star\|^2 + \mathbb{E}[[]\, S_t \sim \mathcal{S}] \frac{1}{|S_t|} \sum_{i \in S_t} \frac{\gamma}{(1+\gamma\mu_i)\mu_i} \|\nabla f_i(x_\star))\|^2 \\
&= A_{\mathcal{S}} \|x_{t-1} - x_\star\|^2 + B_{\mathcal{S}}.
\end{aligned}
$$

where $A_{\mathcal{S}} := \mathbb{E}[[]\, S_t \sim \mathcal{S}] \frac{1}{|S_t|} \sum_{i \in S_t} \frac{1}{1+\gamma\mu_i}$ and $B_{\mathcal{S}} := \mathbb{E}[[]\, S_t \sim \mathcal{S}] \frac{1}{|S_t|} \sum_{i \in S_t} \frac{\gamma}{(1+\gamma\mu_i)\mu_i} \|\nabla f_i(x_\star))\|^2$. By directly applying Fact G.5:

$$
\mathbb{E}[[]\, S_t \sim \mathcal{S}] \|x_t - x_\star\|^2 \leq A_{\mathcal{S}}^t \|x_0 - x_\star\|^2 + \frac{B_{\mathcal{S}}}{1 - A_{\mathcal{S}}}.
$$

□

Lemma G.17 (Inexact formulation of SPPM-AS). *Let $b > 0 \in \mathbb{R}$ and define $\widetilde{\operatorname{prox}}_{\gamma f}(x)$ such that $\forall x \left\| \widetilde{\operatorname{prox}}_{\gamma f}(x) - \operatorname{prox}_{\gamma f}(x) \right\|^2 \leq b$. Let Assumption 2.1 and Assumption 2.2 hold. Let $x_0 \in \mathbb{R}^d$ be an arbitrary starting point. Then for any $t \geq 0$ and any $\gamma > 0$, $s > 0$, the iterates of* SPPM-AS *satisfy*

$$
\mathbb{E}\left[\|x_t - x_\star\|^2 \right] \leq \left(\frac{1+s}{(1+\gamma\mu)^2} \right)^t \|x_0 - x_\star\|^2 + \frac{(1+s)\left(\gamma^2\sigma_\star^2 + s^{-1} b (1+\gamma\mu)^2\right)}{\gamma^2\mu^2 + 2\gamma\mu - s}.
$$

Proof of Lemma G.17. We provide more general version of SPPM proof

$$
\begin{aligned}
\|x_{t+1} - x_\star\|^2 &= \left\|\widetilde{\mathrm{prox}}_{\gamma f_{\xi_t}(x_t)} - \mathrm{prox}_{\gamma f_{\xi_t}}(x_t) + \mathrm{prox}_{\gamma f_{\xi_t}}(x_t) - x_\star\right\|^2 \\
&\overset{Young, s>0}{\leq} (1+s^{-1})\left\|\widetilde{\mathrm{prox}}_{\gamma f_{\xi_t}}(x_t) - \mathrm{prox}_{\gamma f_{\xi_t}}\right\|^2(x_t) + (1+s)\left\|\mathrm{prox}_{\gamma f_{\xi_t}}(x_t) - x_\star\right\|^2 \\
&\leq (1+s^{-1})b + (1+s)\left\|\mathrm{prox}_{\gamma f_{\xi_t}}(x_t) - x_\star\right\|^2.
\end{aligned}
$$

Then proof follows same path as proof Theorem 2.4 and we get

$$
\begin{aligned}
\mathbb{E}\left[\|x_{t+1} - x_\star\|^2\right] &\leq (1+s^{-1})b + (1+s)\frac{1}{(1+\gamma\mu)^2}\left(\|x_t - x_\star\|^2 + \gamma^2\sigma_\star^2\right) \\
&= \frac{1+s}{(1+\gamma\mu)^2}\left(\|x_t - x_\star\|^2 + \left[\gamma^2\sigma_\star^2 + s^{-1}b(1+\gamma\mu)^2\right]\right).
\end{aligned}
$$

azc It only remains to solve the above recursion. Luckily, that is exactly what Fact G.5 does. In particular, we use it with $s_t = \mathbb{E}\left[\|x_t - x_\star\|^2\right]$, $A = \frac{1+s}{(1+\gamma\mu)^2}$ and $B = \frac{(1+s)\left(\gamma^2\sigma_\star^2 + s^{-1}b(1+\gamma\mu)^2\right)}{(1+\gamma\mu)^2}$ to get

$$
\begin{aligned}
\mathbb{E}\left[\|x_t - x_\star\|^2\right] &\leq A^t\|x_0 - x_\star\|^2 + B\frac{1}{1-A} \\
&\leq A^t\|x_0 - x_\star\|^2 + B\frac{(1+\gamma\mu)^2}{(1+\gamma\mu)^2 - 1 - s} \\
&\leq A^t\|x_0 - x_\star\|^2 + \frac{(1+s)\left(\gamma^2\sigma_\star^2 + s^{-1}b(1+\gamma\mu)^2\right)}{(1+\gamma\mu)^2 - 1 - s} \\
&= \left(\frac{1+s}{(1+\gamma\mu)^2}\right)^t\|x_0 - x_\star\|^2 + \frac{(1+s)\left(\gamma^2\sigma_\star^2 + s^{-1}b(1+\gamma\mu)^2\right)}{\gamma^2\mu^2 + 2\gamma\mu - s}.
\end{aligned}
$$

□

G.10 Communication Cost Under Inexact Prox

Proposition G.18 (Communication cost under inexact prox)**.** *Assume the conditions of Theorem 2.4, and let the K-step local prox solver $\mathrm{prox}^A_{\gamma f_S}$ satisfy the mean-squared inexactness bound*

$$
\mathbb{E}\left[\|\,\mathrm{prox}^A_{\gamma f_S}(x) - \mathrm{prox}_{\gamma f_S}(x)\|^2\right] \leq B\rho^K
$$

for some $B > 0$ and $0 < \rho < 1$. Then there exists a stepsize choice $\gamma > 0$ such that the number of global iterations needed to reach $\mathbb{E}\|x_t - x^\star\|^2 \leq \varepsilon$ satisfies

$$
T_\varepsilon \leq \left(\frac{\sigma^2_{\star,\mathrm{AS}} + \rho^K}{2\varepsilon\mu^2_{\mathrm{AS}}} + \frac{1}{2}\right)\log\left(\tfrac{2\|x_0 - x^\star\|^2}{\varepsilon}\right), \tag{23}
$$

for a constant $\kappa > 0$ depending only on the recursion of Lemma G.17. Consequently, the total communication cost (with the hierarchical model) obeys

$$
C_{\mathrm{tot}}(\varepsilon; K) \leq (c_1 K + c_2)\left(\frac{\sigma^2_{\star,\mathrm{AS}} + \rho^K}{2\varepsilon\mu^2_{\mathrm{AS}}} + \frac{1}{2}\right)\log\left(\tfrac{2\|x_0 - x^\star\|^2}{\varepsilon}\right), \tag{24}
$$

so that $C_{\mathrm{tot}}(\varepsilon; K)$ has a finite minimizer $K^\star$ balancing the linear growth of $(c_1 K + c_2)$ and the geometric decay of $B\rho^K$.

Proof. By Lemma G.17, the SPPM-AS iterates with an inexact prox satisfy a recursion of the form

$$
\mathbb{E}\|x_{t+1} - x^\star\|^2 \leq \frac{1}{(1+\gamma\mu_{\mathrm{AS}})^2}\mathbb{E}\|x_t - x^\star\|^2 + \frac{\sigma^2_{\star,\mathrm{AS}} + \kappa B\rho^K}{\mu^2_{\mathrm{AS}}}, \tag{25}
$$

where $\kappa > 0$ depends only on the constants in Lemma G.17. Let

$$u_t := \mathbb{E}\|x_t - x^\star\|^2, \qquad q := \frac{1}{(1+\gamma\mu_{\text{AS}})^2} \in (0,1), \qquad \Sigma_K^2 := \sigma_{\star,\text{AS}}^2 + \kappa B \rho^K.$$

Then equation 25 can be written as

$$u_{t+1} \le q u_t + \frac{\Sigma_K^2}{\mu_{\text{AS}}^2}.$$

Unrolling this linear recursion gives

$$u_t \le q^t u_0 + \frac{\Sigma_K^2}{\mu_{\text{AS}}^2} \sum_{i=0}^{t-1} q^i \le q^t u_0 + \frac{\Sigma_K^2}{\mu_{\text{AS}}^2} \frac{1}{1-q},$$

since $\sum_{i=0}^{t-1} q^i \le 1/(1-q)$ for $q \in (0,1)$.

Next, note that

$$1 - q = 1 - \frac{1}{(1+\gamma\mu_{\text{AS}})^2} = \frac{(1+\gamma\mu_{\text{AS}})^2 - 1}{(1+\gamma\mu_{\text{AS}})^2} \ge \frac{2\gamma\mu_{\text{AS}}}{(1+\gamma\mu_{\text{AS}})^2} \ge \gamma\mu_{\text{AS}},$$

where we used $(1+\gamma\mu_{\text{AS}})^2 - 1 = 2\gamma\mu_{\text{AS}} + \gamma^2\mu_{\text{AS}}^2 \ge 2\gamma\mu_{\text{AS}}$ and $(1+\gamma\mu_{\text{AS}})^2 \le 2(1+\gamma\mu_{\text{AS}})$. Therefore $1/(1-q) \le 1/(\gamma\mu_{\text{AS}})$ and hence

$$u_t \le q^t u_0 + \frac{\Sigma_K^2}{\gamma\mu_{\text{AS}}^3}.$$

As in the iteration-complexity derivation in App. G.5, we now choose

$$\gamma := \frac{\varepsilon\mu_{\text{AS}}}{\Sigma_K^2}$$

and require $\varepsilon \le \Sigma_K^2/\mu_{\text{AS}}^2$ so that the steady-state term is at most $\varepsilon/2$:

$$\frac{\Sigma_K^2}{\gamma\mu_{\text{AS}}^3} = \frac{\Sigma_K^2}{(\varepsilon\mu_{\text{AS}}/\Sigma_K^2)\mu_{\text{AS}}^3} = \frac{\Sigma_K^4}{\varepsilon\mu_{\text{AS}}^4} \le \frac{\varepsilon}{2}.$$

With this choice of γ, the contraction factor satisfies

$$q = \frac{1}{(1+\gamma\mu_{\text{AS}})^2} = \frac{1}{(1+\varepsilon\mu_{\text{AS}}^2/\Sigma_K^2)^2} \le \exp\Big(-2\frac{\varepsilon\mu_{\text{AS}}^2}{\Sigma_K^2}\Big),$$

using $1/(1+z)^2 \le e^{-2z}$ for $z \ge 0$. Thus to ensure the transient term $q^T u_0 \le \varepsilon/2$ it suffices to take

$$T_\varepsilon \ge \frac{1}{2}\Big(1 + \frac{\Sigma_K^2}{\varepsilon\mu_{\text{AS}}^2}\Big) \log\Big(\tfrac{2u_0}{\varepsilon}\Big), \qquad u_0 = \|x_0 - x^\star\|^2,$$

which yields equation 23. Finally, multiplying T_ε by the per-iteration hierarchical communication cost $c_1 K + c_2$ gives equation 24. This proves the proposition. □

G.11 Extensions Beyond Strong Convexity

Where strong convexity is used. Assumption 2.2 (blockwise strong convexity) enters the proof of Theorem 2.4 in exactly two places:

1. *Contractivity of the proximal mapping.* Fact G.4 uses strong convexity of each f_C to conclude that the resolvent $x \mapsto \text{prox}_{\gamma f_C}(x)$ is *strictly contractive* with Lipschitz factor $(1+\gamma\mu_C)^{-1} < 1$. This yields the one-step inequality

$$\|x_{t+1} - x^\star\|^2 \le \frac{1}{(1+\gamma\mu_{\text{AS}})^2} \|x_t - x^\star - \gamma\nabla f_{S_t}(x^\star)\|^2, \tag{26}$$

where μ_{AS} is the aggregate constant from Table 1. Expanding the right-hand side and using $\mathbb{E}[\nabla f_{S_t}(x^\star)] = 0$ produces a linear recurrence in $\mathbb{E}\|x_t - x^\star\|^2$ with noise level $\sigma_{\star,\text{AS}}^2$.

2. *Solving the linear recurrence.* Appendix G.5 solves the scalar recursion

$$\Delta_{t+1} \le \rho(\gamma)\Delta_t + c(\gamma)\sigma^2_{\star,\mathrm{AS}}$$

with $\Delta_t = \mathbb{E}\|x_t - x^\star\|^2$ and $\rho(\gamma) = (1+\gamma\mu_{\mathrm{AS}})^{-2} < 1$, yielding the complexity bound in Theorem 2.4.

Both steps admit standard relaxations beyond strong convexity: (i) quadratic growth / Polyak–Łojasiewicz conditions, which still yield a form of proximal contractivity around the minimizer set; and (ii) weak convexity, where one works instead with the Moreau envelope and a norm of the proximal gradient mapping. We detail these two regimes next.

G.11.1 Quadratic Growth and Proximal PL-type Assumptions

We first record a standard quadratic-growth (QG) condition and show that it suffices to obtain a one-step contraction for the proximal mapping in terms of distance to the minimizer set.

We write $\mathrm{dist}(x, X^\star)$ for the Euclidean distance from x to a set $X^\star \subseteq \mathbb{R}^d$:

$$\mathrm{dist}(x, X^\star) := \inf_{z\in X^\star} \|x - z\|.$$

In the special case $X^\star = \{x^\star\}$, this reduces to $\mathrm{dist}(x, X^\star) = \|x - x^\star\|$.

Definition G.19 (Quadratic growth)**.** *Let $\varphi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be proper, closed, and convex with minimizer set $X^\star := \arg\min_x \varphi(x)$ and optimal value $\varphi^\star$. We say that φ satisfies a* quadratic growth *(QG) condition with constant $\mu_{\mathrm{QG}} > 0$ if*

$$\varphi(x) - \varphi^\star \ge \frac{\mu_{\mathrm{QG}}}{2}\,\mathrm{dist}(x, X^\star)^2 \qquad \forall x \in \mathbb{R}^d. \tag{27}$$

For convex objectives, QG is equivalent to several other conditions such as error bounds and the Polyak–Łojasiewicz (PL) inequality; see, e.g., Karimi et al. (2016) for a detailed comparison.

Lemma G.20 (QG implies contractive prox)**.** *Let φ be proper, closed, and convex, and suppose it satisfies the QG condition equation 27 with constant $\mu_{\mathrm{QG}} > 0$. Fix $\gamma > 0$ and let*

$$x^+ = \mathrm{prox}_{\gamma\varphi}(x) \qquad (\textit{i.e., } x^+ \in \arg\min_z\{\varphi(z) + \tfrac{1}{2\gamma}\|z - x\|^2\}).$$

Then

$$\mathrm{dist}\big(x^+, X^\star\big)^2 \le \frac{1}{1+\gamma\mu_{\mathrm{QG}}}\,\mathrm{dist}\big(x, X^\star\big)^2. \tag{28}$$

In particular, if $X^\star = \{x^\star\}$ is a singleton, then

$$\|x^+ - x^\star\|^2 \le \frac{1}{1+\gamma\mu_{\mathrm{QG}}}\|x - x^\star\|^2. \tag{29}$$

Lemma G.20 shows that the proximal mapping is *strictly contractive in distance to the minimizer set* under QG: the role of strong convexity in Fact G.4 is thus replaced by a QG constant μ_{QG}. If we assume that the global objective f (or each f_C) satisfies a QG/PL condition with constant μ_{PL}, then the proof of Theorem 2.4 can be repeated *verbatim* after substituting Lemma G.20 for Fact G.4, yielding a linear convergence rate

$$\mathbb{E}\big[\mathrm{dist}(x_t, X^\star)^2\big] \le \rho(\gamma)^t\,\mathrm{dist}(x_0, X^\star)^2 + \tilde{c}(\gamma)\sigma^2_{\star,\mathrm{AS}}$$

for some $\rho(\gamma) \in (0,1)$ and a constant $\tilde{c}(\gamma)$ depending on the PL/QG parameter. We do not optimize the constants, since the purpose of this subsection is to clarify that all occurrences of strong convexity in our proof can be replaced by a QG / proximal-PL condition.

G.11.2 Weakly Convex / Non-convex Objectives

We next sketch how SPPM-AS can be interpreted as SGD on a smooth surrogate of a weakly convex objective, which yields the standard sublinear rate to stationarity for a Moreau-envelope-type quantity.

Assumption G.21 (Weak convexity and smoothness)**.** *For each set C with $p_C > 0$, $f_C : \mathbb{R}^d \to \mathbb{R}$ is ρ-weakly convex and L-smooth, i.e.,*

- *$x \mapsto f_C(x) + \frac{\rho}{2}\|x\|^2$ is convex for some $\rho \geq 0$ (independent of C),*
- *∇f_C is L-Lipschitz for some $L \geq 0$ (again independent of C).*

We keep Assumption 2.3 (proper sampling).

Fix $0 < \gamma < 1/\rho$. For each C with $p_C > 0$, define the (componentwise) Moreau envelope

$$f_{C,\gamma}(x) := \min_{z \in \mathbb{R}^d} \bigl\{ f_C(z) + \tfrac{1}{2\gamma}\|z - x\|^2 \bigr\}, \qquad x \in \mathbb{R}^d. \tag{30}$$

Write

$$z_C(x) := \operatorname{prox}_{\gamma f_C}(x) \in \arg\min_z \bigl\{ f_C(z) + \tfrac{1}{2\gamma}\|z - x\|^2 \bigr\},$$

which is single-valued because the objective in equation 30 is strongly convex in z.

It is standard that for ρ-weakly convex f_C and $\gamma < 1/\rho$ the envelope $f_{C,\gamma}$ is continuously differentiable and its gradient can be written in terms of the proximal mapping:

$$\nabla f_{C,\gamma}(x) = \frac{1}{\gamma}\bigl(x - \operatorname{prox}_{\gamma f_C}(x)\bigr) \quad \text{for all } x \in \mathbb{R}^d, \tag{31}$$

and that $\nabla f_{C,\gamma}$ is L_γ-Lipschitz with

$$L_\gamma \leq \max\Bigl\{ \frac{1}{\gamma}, \frac{\rho}{1 - \rho\gamma} \Bigr\}, \tag{32}$$

see, e.g., Lemma 3.1 in Geiersbach & Scarinci (2021). Define the *smoothed global objective*

$$F_\gamma(x) := \mathbb{E}_{S \sim \mathcal{S}}[f_{S,\gamma}(x)]. \tag{33}$$

Lemma G.22 (SPPM-AS as SGD on F_γ)**.** *Under Assumption G.21, with $0 < \gamma < 1/\rho$, let (x_t) be generated by SPPM-AS with exact local proximal steps:*

$$x_{t+1} = \operatorname{prox}_{\gamma f_{S_t}}(x_t), \qquad S_t \sim \mathcal{S} \text{ i.i.d.}$$

Define

$$g_t := \frac{1}{\gamma}\bigl(x_t - x_{t+1}\bigr) = \frac{1}{\gamma}\bigl(x_t - \operatorname{prox}_{\gamma f_{S_t}}(x_t)\bigr). \tag{34}$$

Then:

1. *The iterates admit the "SGD form"*

$$x_{t+1} = x_t - \gamma g_t.$$

2. *F_γ is differentiable with L_γ-Lipschitz gradient for L_γ as in equation 32, and*

$$\mathbb{E}[g_t \mid x_t] = \nabla F_\gamma(x_t). \tag{35}$$

We now give a standard sublinear rate for the norm of the surrogate gradient $\nabla F_\gamma(x_t)$.

Assumption G.23 (Bounded variance)**.** *There exists $\sigma_\gamma^2 \geq 0$ such that for all x,*

$$\mathbb{E}\bigl[\|g_t - \nabla F_\gamma(x_t)\|^2 \mid x_t = x\bigr] \leq \sigma_\gamma^2.$$

Proposition G.24 (Sublinear rate to stationarity of F_γ)**.** *Suppose Assumptions G.21 and G.23 hold and let L_γ be as in equation 32. Run SPPM-AS with step-size $0 < \gamma \leq 1/L_\gamma$, generating (x_t) as in Lemma G.22, and let $T \geq 1$. Then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\bigl[\|\nabla F_\gamma(x_t)\|^2\bigr] \leq \frac{2\bigl(F_\gamma(x_0) - F_\gamma^\star\bigr)}{\gamma T} + L_\gamma \gamma \sigma_\gamma^2, \tag{36}$$

where $F_\gamma^\star := \inf_x F_\gamma(x)$. In particular, taking γ of order $1/\sqrt{T}$ yields the usual $O(1/\sqrt{T})$ decay of the averaged squared gradient norm.

Effect of inexact local solves. We briefly indicate how the "inexact prox" model from Lemma G.17 carries over to the weakly convex regime.

Suppose that instead of the exact update $x_{t+1} = \mathrm{prox}_{\gamma f_{S_t}}(x_t)$ we obtain an approximate proximal point $\tilde{x}_{t+1}$ (for instance, by running K local steps), and define

$$\tilde{g}_t := \frac{1}{\gamma}(x_t - \tilde{x}_{t+1}).$$

Assume that the approximation error satisfies

$$\mathbb{E}\big[\|\tilde{g}_t - g_t\|^2 \mid x_t\big] \le b(K)$$

for some nonincreasing function $b(K) \to 0$ as $K \to \infty$. Then the same argument as in the proof of Proposition G.24, applied with $\tilde{g}_t$ in place of g_t and with variance proxy $\sigma_\gamma^2 + b(K)$, yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla F_\gamma(x_t)\|^2\big] \le \frac{2\big(F_\gamma(x_0) - F_\gamma^\star\big)}{\gamma T} + L_\gamma \gamma\big(\sigma_\gamma^2 + b(K)\big).$$

Thus the only effect of using K local steps instead of an exact proximal oracle is to increase the "noise floor" by a term proportional to $b(K)$, exactly as in the strongly convex analysis with Lemma G.17.

G.12 Within-Round Client Churn (Dropout-Robust SPPM-AS)

Setting. In global round t, the server samples an initial cohort C_t with $|C_t| = C$. During the K intra-cohort sub-iterations $k = 0, \dots, K-1$, some clients may drop; let $C_{t,0} = C_t \supseteq C_{t,1} \supseteq \cdots \supseteq C_{t,K-1}$ be the live sets, and denote $m_k := |C_{t,k}| \ge 1$. Define

$$f_{C_{t,k}}(x) := \frac{1}{m_k}\sum_{i \in C_{t,k}} f_i(x), \qquad P_{t,k}(x) := \mathrm{prox}_{\gamma f_{C_{t,k}}}(x), \qquad P_{t,0}(x) := \mathrm{prox}_{\gamma f_{C_t}}(x).$$

The implemented inner step may be inexact: $\widehat{P}_{t,k}(x) = P_{t,k}(x) + e_{t,k}$.

Assumptions. (i) Each f_i is proper, closed, convex and μ-strongly convex (for some $\mu > 0$); hence any average $f_{C_{t,k}}$ is μ-strongly convex. (ii) For any live set $C_{t,k}$, the arbitrary-sampling constants satisfy $\mu_{\mathrm{AS}}(C_{t,k}) \ge \underline{\mu} > 0$ and $\sigma^2_{\star,\mathrm{AS}}(C_{t,k}) \le \overline{\sigma}^2$. (iii) The inner loop may early-stop when no clients remain, defining $K_{\mathrm{eff}} \le K$ executed sub-iterations.

Assumption G.† (mini-batch subgradient variance). There exists $\sigma^2_{\mathrm{sub}} \ge 0$ such that for any subset $S \subseteq C_t$ with $|S| = m$ and any x,

$$\mathbb{E}\bigg\|\frac{1}{m}\sum_{i\in S} g_i(x) - \frac{1}{C}\sum_{j \in C_t} g_j(x)\bigg\|^2 \le \sigma^2_{\mathrm{sub}}\Big(\frac{1}{m} + \frac{1}{C}\Big), \qquad g_i(x) \in \partial f_i(x),$$

where the expectation is over the sampling mechanism of S (conditional on C_t).

Lemma G.25 (Prox contractivity and composition under shrinking cohorts)**.** *For every k and all x, y,*

$$\|P_{t,k}(x) - P_{t,k}(y)\| \le \frac{1}{1+\gamma\mu}\|x - y\|.$$

Consequently, for the K_{eff}-fold composition,

$$\big\|P_{t,K_{\mathrm{eff}}-1} \circ \cdots \circ P_{t,0}(x) - P_{t,K_{\mathrm{eff}}-1} \circ \cdots \circ P_{t,0}(y)\big\| \le (1+\gamma\mu)^{-K_{\mathrm{eff}}}\|x - y\|.$$

Lemma G.26 (Prox drift between live and initial cohorts)**.** *Fix $x \in \mathbb{R}^d$ and $S \subseteq C_t$ with $|S| = m$, $|C_t| = C$. Let $z_S := \mathrm{prox}_{\gamma f_S}(x)$ and $z_0 := \mathrm{prox}_{\gamma f_{C_t}}(x)$. Then there exists $\tilde{g}_0 \in \partial f_{C_t}(z_S)$ such that for any $g_S \in \partial f_S(z_S)$,*

$$\|z_S - z_0\| \le \frac{\gamma}{1+\gamma\mu}\|g_S - \tilde{g}_0\|.$$

Under Assumption G.†,

$$\mathbb{E}\|z_S - z_0\|^2 \le \Big(\frac{\gamma}{1+\gamma\mu}\Big)^2\Big(\frac{\sigma^2_{\mathrm{sub}}}{m} + \frac{\sigma^2_{\mathrm{sub}}}{C}\Big).$$

Proposition G.27 (Dropout-robust inexact prox; per round, no smoothness)**.** *Let the inner loop in round t execute K_{eff} sub-iterations over $C_{t,0} \supseteq \cdots \supseteq C_{t,K_{\text{eff}}-1}$. Assume μ-strong convexity as above and Assumption G.†. Suppose the local solver errors satisfy $\mathbb{E}\|e_{t,k}\|^2 \leq B\rho^{K_{\text{eff}}}$ for some $B > 0$, $\rho \in (0,1)$. Then*

$$\mathbb{E}\|x_{t+1} - x^\star\|^2 \;\leq\; (1 + \gamma\underline{\mu})^{-2K_{\text{eff}}}\|x_t - x^\star\|^2 \;+\; \frac{\gamma}{\underline{\mu}}\Big(\overline{\sigma}^2 + \kappa B\rho^{K_{\text{eff}}} + \kappa'\Delta_{\text{dr}}(C_{t,\bullet})\Big),$$

for absolute constants $\kappa, \kappa' > 0$, where

$$\Delta_{\text{dr}}(C_{t,\bullet}) \;:=\; \sum_{k=0}^{K_{\text{eff}}-1}\left(\frac{\sigma_{\text{sub}}^2}{m_k} + \frac{\sigma_{\text{sub}}^2}{C}\right) \;=\; K_{\text{eff}}\frac{\sigma_{\text{sub}}^2}{C} \;+\; \sigma_{\text{sub}}^2\sum_{k=0}^{K_{\text{eff}}-1}\frac{1}{m_k}.$$

Corollary G.28 (Elastic-K quorum; explicit bound (no smoothness))**.** *Suppose the inner loop enforces a quorum $\tau \in (0,1]$ (stop if $m_k < \tau C$), so $m_k \geq \tau C$ for all executed sub-iterations and $K_{\text{eff}} \leq K$. Then*

$$\Delta_{\text{dr}}(C_{t,\bullet}) \;\leq\; K_{\text{eff}}\left(\frac{\sigma_{\text{sub}}^2}{C} + \frac{\sigma_{\text{sub}}^2}{\tau C}\right) \;=\; \frac{1+\tau}{\tau}\cdot\frac{K_{\text{eff}}}{C}\sigma_{\text{sub}}^2.$$

Setting

$$\Sigma_K^2 \;:=\; \overline{\sigma}^2 + \kappa B\rho^{K_{\text{eff}}} + \kappa'\frac{1+\tau}{\tau}\cdot\frac{K_{\text{eff}}}{C}\sigma_{\text{sub}}^2,$$

there exists a stepsize choice $\gamma = \varepsilon\underline{\mu}/\Sigma_K^2$ such that

$$T_\varepsilon \;\leq\; \left(\frac{\Sigma_K^2}{2\varepsilon\underline{\mu}^2} + \tfrac{1}{2}\right)\log\left(\tfrac{2\|x_0 - x^\star\|^2}{\varepsilon}\right), \qquad C_{\text{tot}}(\varepsilon; K) \;\leq\; (c_1 K + c_2)T_\varepsilon.$$

Remarks. (i) The analysis uses only μ-strong convexity and subgradient mini-batch variance (Assumption G.†); no smoothness is required. (ii) Higher churn (smaller τ or smaller m_k) enlarges Σ_K^2 and shifts the optimal $K^\star$ downward, but the communication-cost objective still has a finite minimizer. (iii) The early-stop rule (quorum τ) is a practical safeguard that ensures bounded drift and keeps the theory in-range under churn.

We verify our theory through extensive experiments. We first calculate the executed subrounds under within-round churn in Figure 16. Here s denotes the expected fraction of the selected cohort that participates in each intra-cohort synchronization, and the quorum τ means the inner loop proceeds only while the live set size satisfies $m_k \geq \tau C$. This measurement makes explicit how within-round churn limits the number of inner synchronizations that are actually executed. With active ratio s, the live cohort in each subround is about $m_k \approx sC$, and the inner loop proceeds only while $m_k \geq \tau C$ (quorum τ). Hence $K_{\text{eff}} < K$ and, as the figure shows, it saturates well below K : around 2.6 when $s = 0.8$ and around 5 when $s = 0.9$, largely independent of the nominal K. This directly aligns with our dropout-robust recursion in App. $\sim$ G, where one round depends on K_{eff} (not K): the inexact-prox term shrinks geometrically as $\rho^{K_{\text{eff}}}$ while the churn penalty grows only linearly with K_{eff}/C. Consequently, moderate churn above the quorum ($s = 0.8$ with $\tau = 0.7$) does not materially change the final neighborhood reported elsewhere, because K_{eff} is already large enough that $\rho^{K_{\text{eff}}}$ is tiny; what differs is smoothness-higher s yields larger m_k and therefore lower per-sync gradient variance, producing slightly steadier traces. The saturation also explains why increasing the nominal K under churn may add communication without yielding additional executed subrounds, which is precisely the effect captured by our communication-cost bound $C_{\text{tot}}(\varepsilon; K)$.

We alse vary the within-round active ratio s, which controls the expected live set $m_k \approx sC$ at each intra-cohort synchronization, while enforcing a quorum $m_k \geq \tau C$ with $\tau = 0.7$ (Figure 17). Our dropout-robust recursion predicts that the steady-state neighborhood depends on $\Sigma_K^2 \propto B\rho^{K_{\text{eff}}} + \Delta_{\text{dr}}$, where $\Delta_{\text{dr}} = \sum_{k=0}^{K_{\text{eff}}-1}\big(\sigma_{\text{sub}}^2/m_k + \sigma_{\text{sub}}^2/C\big)$. Under the quorum, $m_k \geq \tau C$ implies $\Delta_{\text{dr}} \leq \frac{1+\tau}{\tau}\frac{K_{\text{eff}}}{C}\sigma_{\text{sub}}^2$, which depends on τ but only weakly on s as long as $s \geq \tau$. This is exactly what we observe: $s = 0.8$ produces essentially the same final neighborhood as $s = 1.0$. The slightly smoother curves at $s = 1.0$ follow from the variance term $\sigma_{\text{sub}}^2/m_k$, which is smaller when more clients are active. For the per-sync learning rate, the theoretical choice that maximizes a safe contraction under expected-smoothness with a live set of size m is

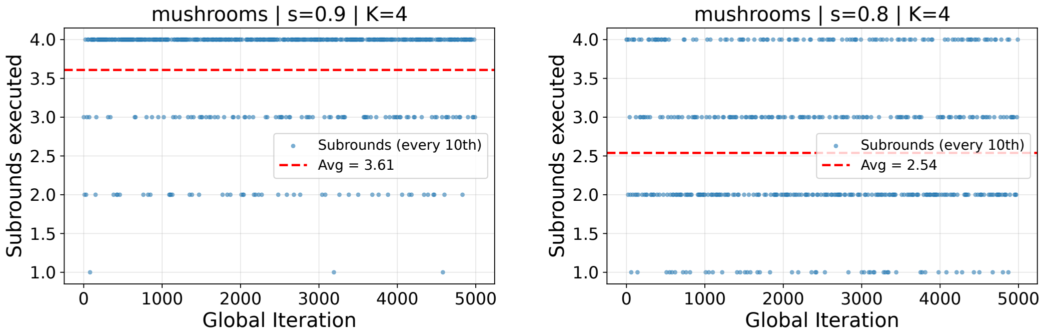


Figure 16: Executed subrounds under within-round churn on `mushrooms` (quorum $\tau = 0.7$). Dots show the number of intra-cohort synchronizations completed in each global iteration (sampled every 10th); the dashed line is the mean K_{eff}. For s=0.9, $K_{\text{eff}} \approx 3.61$ while for s=0.8, $K_{\text{eff}} \approx 2.54$.

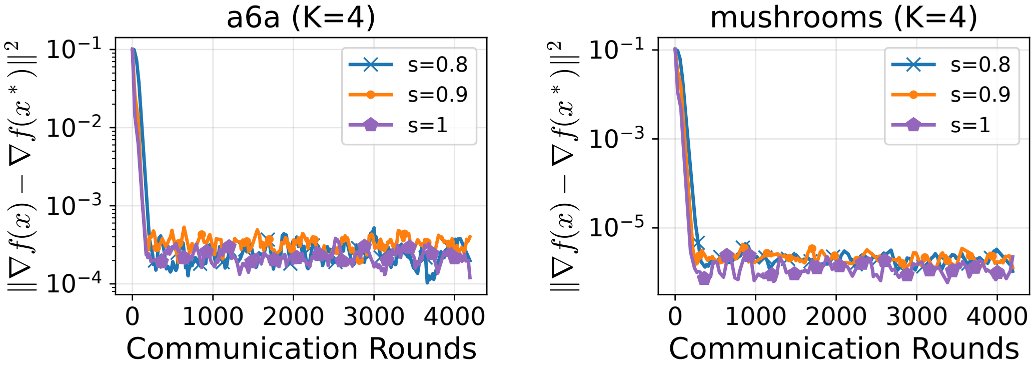


Figure 17: Effect of within-round active ratio s (quorum $\tau = 0.7$, $K = 4$). Lowering s from 1.0 to 0.8 leaves the final stationarity essentially unchanged on both datasets, while s=1.0 yields slightly smoother trajectories.

$$\eta^\star(m) = \frac{1}{\widetilde{L}_m}, \qquad \widetilde{L}_m = \frac{(n-m)L_{\max} + n(m-1)L_{\text{avg}}}{n-1},$$

which reduces to $\eta^\star(m) \approx 1/(mL)$ when $L_i \approx L$. In practice, to keep updates stable under churn and to equalize progress across different s, we use the conservative rule

$$\eta_{\text{practical}}(m) = \eta_0 \times \frac{m}{C} = \eta_0 \times s,$$

where η_0 is tuned at full participation $m = C$. Since $\eta_{\text{practical}}(m) \le \eta_0 \le 1/\widetilde{L}_C \le 1/\widetilde{L}_m$ for all $m \le C$, this schedule is always within the theoretical stability range and explains why $s = 0.8$ is only slightly noisier yet attains the same neighborhood as $s = 1.0$. Overall, the figure confirms that the method is robust to moderate within-round churn above the quorum (unchanged neighborhood), while higher s primarily improves smoothness.

G.13 FedAvg-K-Reuse Baseline: Algorithm and Theory

Algorithm (definition). In each global round t, the server samples a cohort $C_t \subseteq [n]$ of size C once, and then performs $K \ge 1$ *intra-cohort synchronizations* with the same cohort. **K-Reuse communicates after every local gradient step**:

$$x_{t,0} = x_t, \qquad x_{t,k+1} = x_{t,k} - \eta \frac{1}{C} \sum_{i \in C_t} \nabla f_i(x_{t,k}), \quad k = 0, \ldots, K-1, \qquad x_{t+1} = x_{t,K}.$$

This is one gradient step per client followed by averaging, repeated K times with no cohort resampling inside the round.

Assumptions.

(A1) (**Strong convexity**) Each f_i is μ-strongly convex for a common $\mu > 0$. Hence $f_C(x) := \frac{1}{|C|}\sum_{i\in C} f_i(x)$ is μ-strongly convex for any cohort C. Denote $x^\star := \arg\min_x f(x)$ with $f := \frac{1}{n}\sum_{i=1}^n f_i$.

(A2) (**Lipschitz gradients**) Each f_i is L_i-smooth. Define $L_{\max} := \max_i L_i$ and $L_{\text{avg}} := \frac{1}{n}\sum_{i=1}^n L_i$. For a uniformly sampled cohort C_t of size C, the cohort average f_{C_t} is L_{C_t}-smooth with $L_{C_t} = \frac{1}{C}\sum_{i\in C_t} L_i \le L_{\max}$.

(A3) (**Cohort-gradient variance at the optimum**) Let

$$\sigma_\star^2 \;:=\; \mathbb{E}_C \big\|\nabla f_C(x^\star)\big\|^2,$$

where C is a uniformly sampled size-C subset of $[n]$.

Per-synchronization stepsize (communication every local step). Let n be the total number of clients and C the cohort size. The *expected-smoothness* constant for uniform mini-batches of size C (without replacement) is

$$\widetilde{L}_C \;:=\; \frac{(n-C)L_{\max} + n(C-1)L_{\text{avg}}}{n-1}.$$

We choose the per-synchronization stepsize

$$\boxed{\eta \;=\; \frac{1}{\widetilde{L}_C} \;=\; \frac{n-1}{(n-C)L_{\max} + n(C-1)L_{\text{avg}}}} \tag{37}$$

which is exactly the simplified closed form of the `LocalGD` stepsize (Khaled et al., 2019) when one gradient is taken per synchronization. Since K-Reuse communicates every local step, no additional scaling by a local step counter is needed.

We now quantify the within-round contraction and the sampling bias from optimizing f_{C_t} instead of f.

Lemma G.29 (GD contraction on a fixed cohort)**.** *Fix a cohort C and suppose f_C is μ-strongly convex and L_C-smooth. For any $\eta \in (0, 2/(\mu + L_C)]$, one GD step on f_C satisfies*

$$\|x^+ - x_C^\star\|^2 \;\le\; \rho_C^2(\eta)|x - x_C^\star\|^2, \qquad \rho_C(\eta) := \max\{|1-\eta\mu|,\ |1-\eta L_C|\} \;\le\; \frac{L_C-\mu}{L_C+\mu}.$$

Consequently, K synchronizations with the same cohort yield

$$\|x_{t,K} - x_C^\star\|^2 \;\le\; \rho_C^{2K}(\eta)\|x_t - x_C^\star\|^2, \qquad \rho_C(\eta) \le 1-\eta\mu.$$

Lemma G.30 (Cohort minimizer vs. global minimizer)**.** *Let $x_C^\star := \arg\min f_C$ and $x^\star := \arg\min f$. Then*

$$\|x_C^\star - x^\star\| \;\le\; \frac{1}{\mu}\|\nabla f_C(x^\star)\|.$$

In particular, $\mathbb{E}\|x_C^\star - x^\star\|^2 \le \sigma_\star^2/\mu^2$.

Proposition G.31 (Round-wise recursion for FedAvg-K-reuse)**.** *Let x_{t+1} be produced by K synchronizations with cohort C_t and stepsize η satisfying $0 < \eta \le 2/(\mu + L_{C_t})$. Then*

$$\|x_{t+1} - x^\star\|^2 \;\le\; \rho_{C_t}^{2K}(\eta)\|x_t - x^\star\|^2 \;+\; 2(1+\rho_{C_t}^{2K}(\eta))\frac{\|\nabla f_{C_t}(x^\star)\|^2}{\mu^2}.$$

Taking expectations over the random cohort,

$$\mathbb{E}\|x_{t+1} - x^\star\|^2 \;\le\; (1-\eta\mu)^{2K}\|x_t - x^\star\|^2 \;+\; \frac{2\sigma_\star^2}{\mu^2}.$$

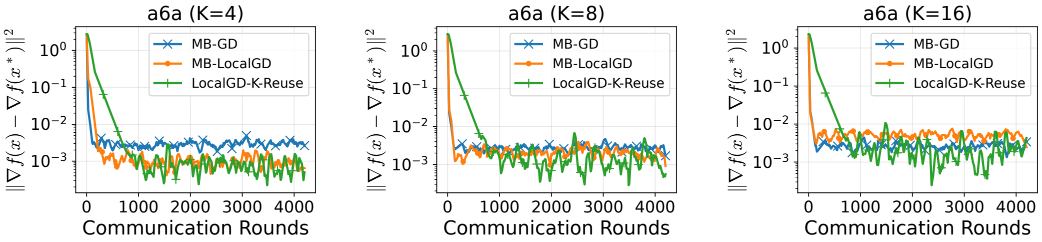


Figure 18: MB-GD vs. MB-LocalGD vs. LocalGD-K-Reuse on a6a. LocalGD-K-Reuse achieves lower error per communication but shows larger oscillations; MB-LocalGD degrades at large K due to client drift.

Discussion. (i) **Communication every local step.** In K-Reuse, each local gradient step is followed by a synchronization, so a round with K steps incurs K communications; the per-sync stepsize in equation 37 is calibrated for one gradient per sync and does not depend on K. (ii) **Ablation against SPPM-AS.** With the same cohort and the same K, K-Reuse isolates the effect of cohort reuse without proximal anchoring.

In Figure 18, we show the comparison among MB-GD, MB-LocalGD and LocalGD-K-Reuse. Across all panels, LocalGD-K-Reuse eventually reaches a lower $\|\nabla f(x) - \nabla f(x^\star)\|^2$ than MB-GD and MB-LocalGD, even though its early decrease per communication can be slower and visibly more oscillatory. This behavior is consistent with the theory: by synchronizing every local step, K-Reuse eliminates the K-dependent client-drift term that inflates the asymptotic neighborhood of MB-LocalGD on non-IID data, so its steady-state error floor is smaller; however, each communication in K-Reuse corresponds to a single cohort GD step, whereas MB-LocalGD effectively accumulates K local steps before averaging, which can produce a steeper initial slope. The oscillations arise because repeated *unanchored* steps pull toward the current cohort minimizer $x^\star_{C_t}$, and the target shifts when the cohort changes; equalizing the per-round step ($\eta \leftarrow \eta/K$) or adding a mild proximal anchor reduces this ringing. At $K{=}16$, the heterogeneity-driven drift accumulated by MB-LocalGD becomes large enough that it underperforms even MB-GD, while K-Reuse remains robust thanks to frequent synchronization.

In Figure 19, we also compare LocalGD-K-Reuse with our core algorithm SPPM-AS. SPPM-AS consistently achieves a smaller neighborhood with smoother trajectories than LocalGD-K-Reuse because each update approximates $x_{t+1} \approx \mathrm{prox}\,\gamma f C_t(x_t)$, a firmly non-expansive mapping that contracts by $(1+\gamma\mu_{\mathrm{AS}})^{-2}$ and damps cohort-switching. The inexact-prox bias from the K-round local solver decays geometrically as $B\rho^K$, so additional intra-cohort rounds reduce error without compounding drift. By contrast, K-Reuse is explicit GD on f_{C_t}; with the same step size used at $K{=}1$ or under changing live-set sizes, composing K steps per round can be under-damped and visibly oscillatory.

G.14 G.11-G.13 Missing Proofs

G.14.1 Appendix G.11 Proofs

Proof of Lemma G.20

Proof. Let $x^+ = \mathrm{prox}_{\gamma\varphi}(x)$. By optimality of x^+, there exists a subgradient $g^+ \in \partial\varphi(x^+)$ such that

$$0 \in g^+ + \frac{1}{\gamma}(x^+ - x) \quad \Longleftrightarrow \quad g^+ = \frac{1}{\gamma}(x - x^+). \tag{38}$$

Let $x^\star_+ \in X^\star$ be a (Euclidean) projection of x^+ onto $X^\star$, i.e.,

$$x^\star_+ \in \arg\min_{z \in X^\star} \|x^+ - z\| \qquad \text{so that} \qquad \|x^+ - x^\star_+\| = \mathrm{dist}(x^+, X^\star).$$

Step 1: lower bound via QG. By Definition G.19 with $x = x^+$ and $X^\star$,

$$\varphi(x^+) - \varphi^\star \geq \frac{\mu_{\mathrm{QG}}}{2}\|x^+ - x^\star_+\|^2. \tag{39}$$

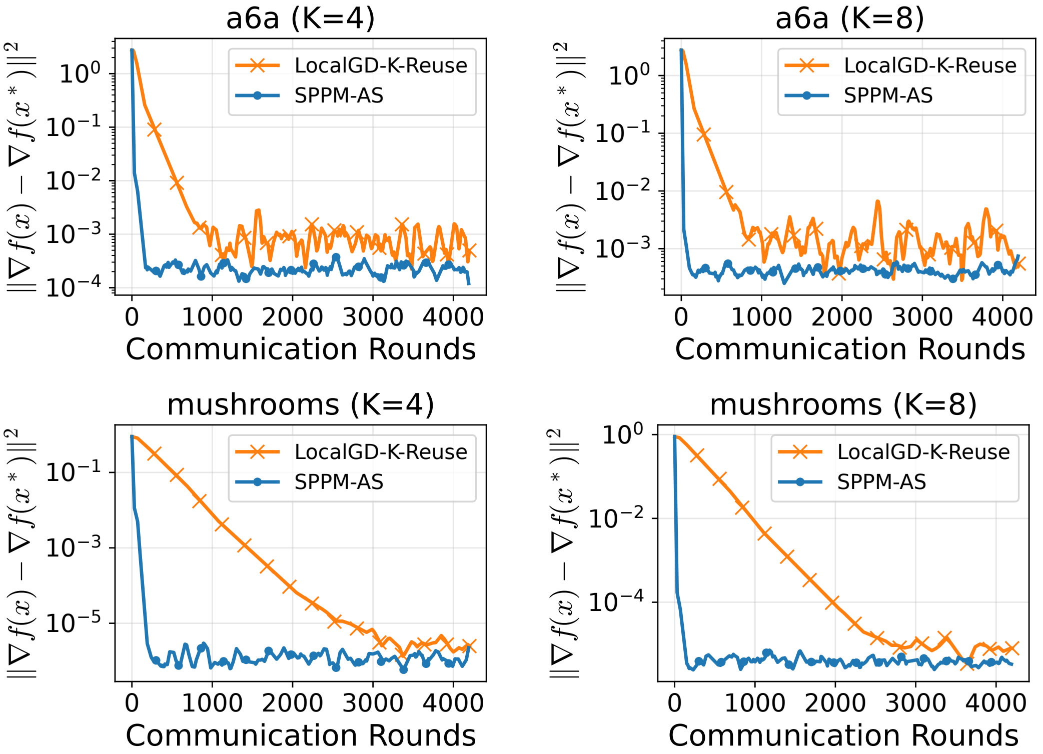


Figure 19: SPPM-AS vs. LocalGD-K-Reuse on a6a and mushrooms. SPPM-AS is faster and markedly more stable; the proximal anchor damps cohort-switching oscillations.

Step 2: upper bound via convexity and prox optimality. By convexity of φ, for any y and any $g \in \partial\varphi(y)$ we have

$$\varphi(z) \geq \varphi(y) + \langle g, z - y \rangle \quad \forall z.$$

Taking $y = x^+$, $z = x_+^\star$, and $g = g^+$, and noting that $x_+^\star$ is a minimizer so $\varphi(x_+^\star) = \varphi^\star$, we obtain

$$\varphi^\star \geq \varphi(x^+) + \langle g^+, x_+^\star - x^+ \rangle = \varphi(x^+) - \langle g^+, x^+ - x_+^\star \rangle.$$

Rearranging and inserting equation 38,

$$\varphi(x^+) - \varphi^\star \leq \langle g^+, x^+ - x_+^\star \rangle = \frac{1}{\gamma}\langle x - x^+, x^+ - x_+^\star \rangle. \tag{40}$$

Step 3: combine the bounds. Combining equation 39 and equation 40 gives

$$\frac{\mu_{\mathrm{QG}}}{2}\|x^+ - x_+^\star\|^2 \leq \frac{1}{\gamma}\langle x - x^+, x^+ - x_+^\star \rangle.$$

Multiplying by γ and rearranging,

$$\langle x - x^+, x^+ - x_+^\star \rangle \geq \frac{\gamma\mu_{\mathrm{QG}}}{2}\|x^+ - x_+^\star\|^2. \tag{41}$$

Step 4: expand $\|x - x_+^\star\|^2$. Using the parallelogram identity,

$$\|x - x_+^\star\|^2 = \|x - x^+\|^2 + 2\langle x - x^+, x^+ - x_+^\star \rangle + \|x^+ - x_+^\star\|^2.$$

Applying equation 41,

$$\|x - x_+^\star\|^2 \geq \|x - x^+\|^2 + \gamma\mu_{\mathrm{QG}}\|x^+ - x_+^\star\|^2 + \|x^+ - x_+^\star\|^2 = \|x - x^+\|^2 + (1 + \gamma\mu_{\mathrm{QG}})\|x^+ - x_+^\star\|^2.$$

Dropping the nonnegative term $\|x - x^+\|^2$ yields

$$(1 + \gamma\mu_{\mathrm{QG}})\|x^+ - x_+^\star\|^2 \leq \|x - x_+^\star\|^2.$$

Step 5: relate to $\operatorname{dist}(x, X^\star)$. Since $x_+^\star \in X^\star$ is an arbitrary element of the minimizer set (specifically, a projection of x^+ onto $X^\star$), we have

$$\operatorname{dist}(x, X^\star)^2 = \min_{z \in X^\star} \|x - z\|^2 \le \|x - x_+^\star\|^2.$$

Combining with the previous inequality,

$$(1 + \gamma\mu_{\mathrm{QG}}) \operatorname{dist}(x^+, X^\star)^2 = (1 + \gamma\mu_{\mathrm{QG}})\|x^+ - x_+^\star\|^2 \le \|x - x_+^\star\|^2 \ge \operatorname{dist}(x, X^\star)^2,$$

which gives equation 28. If $X^\star = \{x^\star\}$ is a singleton, then for all x, $\operatorname{dist}(x, X^\star) = \|x - x^\star\|$, and hence equation 29 follows.

□

Proof of Lemma G.22

Proof. The identity $x_{t+1} = x_t - \gamma g_t$ follows immediately from the definition equation 34. By equation 31,

$$\nabla f_{S_t,\gamma}(x_t) = \frac{1}{\gamma}\big(x_t - \operatorname{prox}_{\gamma f_{S_t}}(x_t)\big) = g_t.$$

Taking conditional expectation with respect to S_t given x_t and using definition equation 33,

$$\mathbb{E}[g_t \mid x_t] = \mathbb{E}_{S_t}\big[\nabla f_{S_t,\gamma}(x_t) \mid x_t\big] = \nabla F_\gamma(x_t),$$

which is equation 35. Finally, F_γ is differentiable with L_γ-Lipschitz gradient because it is an expectation of functions $f_{C,\gamma}$ that each have L_γ-Lipschitz gradients, see equation 32. □

Proof of Proposition G.24

Proof. Since F_γ is differentiable with L_γ-Lipschitz gradient, it satisfies the standard smoothness inequality:

$$F_\gamma(x_{t+1}) \le F_\gamma(x_t) + \big\langle \nabla F_\gamma(x_t), x_{t+1} - x_t \big\rangle + \frac{L_\gamma}{2}\|x_{t+1} - x_t\|^2.$$

Using $x_{t+1} = x_t - \gamma g_t$ from Lemma G.22,

$$F_\gamma(x_{t+1}) \le F_\gamma(x_t) - \gamma\big\langle \nabla F_\gamma(x_t), g_t \big\rangle + \frac{L_\gamma \gamma^2}{2}\|g_t\|^2.$$

Take conditional expectation with respect to x_t. By equation 35,

$$\mathbb{E}\big[F_\gamma(x_{t+1}) \mid x_t\big] \le F_\gamma(x_t) - \gamma\|\nabla F_\gamma(x_t)\|^2 + \frac{L_\gamma \gamma^2}{2}\mathbb{E}\big[\|g_t\|^2 \mid x_t\big].$$

Note that

$$\mathbb{E}\big[\|g_t\|^2 \mid x_t\big] = \mathbb{E}\big[\|g_t - \nabla F_\gamma(x_t) + \nabla F_\gamma(x_t)\|^2 \mid x_t\big] \le 2\|\nabla F_\gamma(x_t)\|^2 + 2\mathbb{E}\big[\|g_t - \nabla F_\gamma(x_t)\|^2 \mid x_t\big],$$

so by Assumption G.23,

$$\mathbb{E}\big[\|g_t\|^2 \mid x_t\big] \le 2\|\nabla F_\gamma(x_t)\|^2 + 2\sigma_\gamma^2.$$

Substituting,

$$\mathbb{E}\big[F_\gamma(x_{t+1}) \mid x_t\big] \le F_\gamma(x_t) - \gamma\|\nabla F_\gamma(x_t)\|^2 + L_\gamma\gamma^2\|\nabla F_\gamma(x_t)\|^2 + L_\gamma\gamma^2\sigma_\gamma^2.$$

Since $\gamma \le 1/L_\gamma$, we have $1 - L_\gamma\gamma \ge 0$ and thus

$$\mathbb{E}\big[F_\gamma(x_{t+1}) \mid x_t\big] \le F_\gamma(x_t) - \frac{\gamma}{2}\|\nabla F_\gamma(x_t)\|^2 + L_\gamma\gamma^2\sigma_\gamma^2.$$

Taking expectation over x_t, rearranging, and using $F_\gamma(x_{t+1}) \ge F_\gamma^\star$,

$$\frac{\gamma}{2}\mathbb{E}\big[\|\nabla F_\gamma(x_t)\|^2\big] \le \mathbb{E}\big[F_\gamma(x_t) - F_\gamma(x_{t+1})\big] + L_\gamma\gamma^2\sigma_\gamma^2.$$

Summing over $t = 0, \dots, T-1$,

$$\frac{\gamma}{2}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla F_\gamma(x_t)\|^2\big] \le F_\gamma(x_0) - F_\gamma(x_T) + L_\gamma\gamma^2 T\sigma_\gamma^2 \le F_\gamma(x_0) - F_\gamma^\star + L_\gamma\gamma^2 T\sigma_\gamma^2.$$

Dividing by $\gamma T/2$ yields equation 36. □

G.14.2 Appendix G.12 Proofs

Proof of Lemma G.25

Proof. Fix k and set $T := P_{t,k} = \mathrm{prox}_{\gamma f_{C_{t,k}}}$. Let $u = T(x)$ and $v = T(y)$. By prox optimality, there exist $s_u \in \partial f_{C_{t,k}}(u)$ and $s_v \in \partial f_{C_{t,k}}(v)$ such that

$$\frac{1}{\gamma}(x-u) \in \partial f_{C_{t,k}}(u), \qquad \frac{1}{\gamma}(y-v) \in \partial f_{C_{t,k}}(v),$$

i.e., $x - u = \gamma s_u$ and $y - v = \gamma s_v$. Subtracting gives

$$u - v \;=\; (x-y) - \gamma(s_u - s_v).$$

Taking the inner product with $u - v$ and using Cauchy–Schwarz,

$$\|u-v\|^2 \;=\; \langle x-y, u-v\rangle - \gamma\langle s_u - s_v, u-v\rangle \;\le\; \|x-y\|\|u-v\| - \gamma\langle s_u - s_v, u-v\rangle.$$

Since $f_{C_{t,k}}$ is μ-strongly convex, its subdifferential is μ-strongly monotone: $\langle s_u - s_v, u-v\rangle \ge \mu\|u-v\|^2$. Hence

$$\|u-v\|^2 \;\le\; \|x-y\|\|u-v\| - \gamma\mu\|u-v\|^2 \;\;\Longrightarrow\;\; (1+\gamma\mu)\|u-v\| \;\le\; \|x-y\|.$$

Thus $\|T(x) - T(y)\| \;\le\; (1+\gamma\mu)^{-1}\|x-y\|$. The composition bound follows by multiplying Lipschitz constants across the K_{eff} maps. □

Proof of Lemma G.26

Proof. Let $g_S \in \partial f_S(z_S)$ be the subgradient that certifies the prox optimality $x - z_S = \gamma g_S$. Since f_{C_t} is convex, $\partial f_{C_t}(z_S) \neq \emptyset$. Pick any selection $\tilde g_0 \in \partial f_{C_t}(z_S)$. Also let $g_0 \in \partial f_{C_t}(z_0)$ satisfy $x - z_0 = \gamma g_0$. Subtract the two optimality conditions: $z_S - z_0 = \gamma\big((g_0 - \tilde g_0) + (\tilde g_0 - g_S)\big)$. Take inner product with $z_S - z_0$:

$$\|z_S - z_0\|^2 = \gamma\langle g_0 - \tilde g_0, z_S - z_0\rangle \;+\; \gamma\langle \tilde g_0 - g_S, z_S - z_0\rangle.$$

By μ-strong monotonicity of ∂f_{C_t} at the pair (z_S, z_0), $\langle g_0 - \tilde g_0, z_S - z_0\rangle \ge \mu\|z_S - z_0\|^2$. Hence

$$\|z_S - z_0\|^2 \;\le\; \gamma\|\tilde g_0 - g_S\|\|z_S - z_0\| - \gamma\mu\|z_S - z_0\|^2,$$

which yields $(1+\gamma\mu)\|z_S - z_0\| \le \gamma\|\tilde g_0 - g_S\|$ and the first claim: $\|z_S - z_0\| \le \frac{\gamma}{1+\gamma\mu}\|\tilde g_0 - g_S\|$.

For the variance bound, choose the selections so that $g_S(z_S) = \frac{1}{m}\sum_{i\in S} g_i(z_S)$ and $\tilde g_0(z_S) = \frac{1}{C}\sum_{j\in C_t} g_j(z_S)$ with $g_i(z_S) \in \partial f_i(z_S)$. Assumption G.† at the common point z_S gives $\mathbb{E}\|g_S(z_S) - \tilde g_0(z_S)\|^2 \le \sigma_{\text{sub}}^2(\frac{1}{m} + \frac{1}{C})$, whence the stated inequality by multiplying with $\big(\frac{\gamma}{1+\gamma\mu}\big)^2$. □

Proof of Proposition G.27

Proof. Let $\alpha := \frac{1}{1+\gamma\mu} < 1$. Define the *exact* inner sequence $y_{t,0} := x_t$ and $y_{t,k+1} := P_{t,k}(y_{t,k})$ for $k = 0, \ldots, K_{\text{eff}} - 1$; and the *implemented* inner sequence $x_{t,0} := x_t$ and $x_{t,k+1} := P_{t,k}(x_{t,k}) + e_{t,k}$.

Step 1: algorithmic inexactness. Let $\delta_k := x_{t,k} - y_{t,k}$. By Lemma G.25,

$$\|\delta_{k+1}\| = \|P_{t,k}(x_{t,k}) - P_{t,k}(y_{t,k}) + e_{t,k}\| \;\le\; \alpha\|\delta_k\| + \|e_{t,k}\|.$$

Unrolling from $\delta_0 = 0$,

$$\|\delta_{K_{\text{eff}}}\| \;\le\; \sum_{j=0}^{K_{\text{eff}}-1} \alpha^{K_{\text{eff}}-1-j}\|e_{t,j}\|.$$

By Cauchy–Schwarz,

$$\|\delta_{K_{\text{eff}}}\|^2 \;\le\; \Big(\sum_{j=0}^{K_{\text{eff}}-1} \alpha^{2(K_{\text{eff}}-1-j)}\Big)\Big(\sum_{j=0}^{K_{\text{eff}}-1} \|e_{t,j}\|^2\Big) \;\le\; \frac{1}{1-\alpha^2}\sum_{j=0}^{K_{\text{eff}}-1} \|e_{t,j}\|^2.$$

Taking expectations and using $\mathbb{E}\|e_{t,j}\|^2 \le B\rho^{K_{\text{eff}}}$ yields

$$\mathbb{E}\|x_{t,K_{\text{eff}}} - y_{t,K_{\text{eff}}}\|^2 \;\le\; \kappa_1 B\rho^{K_{\text{eff}}}, \qquad \kappa_1 := \frac{K_{\text{eff}}}{1-\alpha^2}.$$

Step 2: operator drift within the round. Introduce a *reference* inner sequence driven by the initial operator $P_{t,0}$ but evaluated along the exact inputs: $z_{t,0} := x_t$, $z_{t,k+1} := P_{t,0}(y_{t,k})$. Then

$$\|y_{t,k+1} - z_{t,k+1}\| \;=\; \|P_{t,k}(y_{t,k}) - P_{t,0}(y_{t,k})\|.$$

Define $r_k := z_{t,k} - P_{t,0}(z_{t,k-1})$ for $k \ge 1$ so that r_k captures how the input of $P_{t,0}$ deviates from its own trajectory. Using α-Lipschitzness of $P_{t,0}$,

$$\|z_{t,k} - P_{t,0}(z_{t,k-1})\| \;=\; \|P_{t,0}(y_{t,k-1}) - P_{t,0}(z_{t,k-1})\| \;\le\; \alpha\|y_{t,k-1} - z_{t,k-1}\|.$$

A standard perturbation argument for compositions of contractive maps (prove by induction) gives

$$\|y_{t,K_{\text{eff}}} - z_{t,K_{\text{eff}}}\| \;\le\; \sum_{j=0}^{K_{\text{eff}}-1} \alpha^{K_{\text{eff}}-1-j}\|P_{t,j}(y_{t,j}) - P_{t,0}(y_{t,j})\|.$$

Apply Lemma G.26 at the points $x = y_{t,j}$ and take expectations:

$$\mathbb{E}\|P_{t,j}(y_{t,j}) - P_{t,0}(y_{t,j})\|^2 \;\le\; \left(\frac{\gamma}{1+\gamma\underline{\mu}}\right)^2\left(\frac{\sigma_{\text{sub}}^2}{m_j} + \frac{\sigma_{\text{sub}}^2}{C}\right).$$

Cauchy–Schwarz yields

$$\mathbb{E}\|y_{t,K_{\text{eff}}} - z_{t,K_{\text{eff}}}\|^2 \;\le\; \frac{1}{1-\alpha^2}\left(\frac{\gamma}{1+\gamma\underline{\mu}}\right)^2 \sum_{j=0}^{K_{\text{eff}}-1}\left(\frac{\sigma_{\text{sub}}^2}{m_j} + \frac{\sigma_{\text{sub}}^2}{C}\right) \;=\; \kappa_2\Delta_{\text{dr}}(C_{t,\bullet}),$$

for $\kappa_2 := \frac{1}{1-\alpha^2}\left(\frac{\gamma}{1+\gamma\underline{\mu}}\right)^2$.

Step 3: contraction to the global optimum and aggregation. We decompose

$$\|x_{t+1} - x^\star\| \;\le\; \|x_{t+1} - y_{t,K_{\text{eff}}}\| + \|y_{t,K_{\text{eff}}} - z_{t,K_{\text{eff}}}\| + \|z_{t,K_{\text{eff}}} - x^\star\|.$$

Squaring and using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, then taking expectations, and plugging the bounds from Steps 1–2 gives

$$\mathbb{E}\|x_{t+1} - x^\star\|^2 \;\le\; 3\kappa_1 B\rho^{K_{\text{eff}}} \;+\; 3\kappa_2\Delta_{\text{dr}}(C_{t,\bullet}) \;+\; 3\mathbb{E}\|z_{t,K_{\text{eff}}} - x^\star\|^2.$$

It remains to bound the last term. Note $z_{t,k+1} = P_{t,0}(y_{t,k})$ and $P_{t,0}$ is α-Lipschitz around $x^\star$ with arbitrary-sampling neighborhood bounded by $\overline{\sigma}^2$ by Assumption (ii). The exact SPPM recursion (the main theorem for exact prox with AS constants) implies for one step

$$\mathbb{E}\|P_{t,0}(u) - x^\star\|^2 \;\le\; \alpha^2\|u - x^\star\|^2 + \frac{\gamma}{\underline{\mu}}\overline{\sigma}^2,$$

hence iterating K_{eff} times and upper-bounding the geometric series by $\frac{1}{1-\alpha^2} \le \frac{1+\gamma\underline{\mu}}{2\gamma\underline{\mu}} \le \frac{1}{\gamma\underline{\mu}}$ gives

$$\mathbb{E}\|z_{t,K_{\text{eff}}} - x^\star\|^2 \;\le\; \alpha^{2K_{\text{eff}}}\|x_t - x^\star\|^2 \;+\; \frac{\gamma}{\underline{\mu}}\overline{\sigma}^2.$$

Collect constants (absorbing the factor 3 into κ, κ') to obtain the stated inequality. □

Proof of Corollary G.28

Proof. From $m_k \ge \tau C$ we have $\sum_{k=0}^{K_{\text{eff}}-1}(1/m_k) \le K_{\text{eff}}/(\tau C)$, which gives the bound on $\Delta_{\text{dr}}(C_{t,\bullet})$ directly. Substitute this in Proposition G.27, set $\alpha = \frac{1}{1+\gamma\underline{\mu}}$, and follow the same iteration-complexity derivation as in the main text (choosing γ to minimize the usual upper bound) to obtain the displayed T_ε and hence the communication-cost bound. □

G.14.3 Appendix G.13 Proofs

Proof of Lemma G.29

Proof. Let $f = f_C$ and $x^\star = x_C^\star$. For L_C-smooth f, the descent lemma gives

$$f(x-\eta\nabla f(x)) \;\le\; f(x) - \eta\Big(1-\frac{\eta L_C}{2}\Big)\|\nabla f(x)\|^2 \qquad \text{for } \eta\in(0,2/L_C].$$

For μ-strongly convex f, $\|\nabla f(x)\|^2 \ge 2\mu\big(f(x)-f(x^\star)\big)$. Combine the two to obtain

$$f(x^+) - f(x^\star) \;\le\; \big(1-\eta\mu\big)\big(f(x)-f(x^\star)\big) \qquad (\eta \le 1/L_C).$$

Strong convexity also implies $2\mu\|x-x^\star\|^2 \le \frac{2}{\mu}\|\nabla f(x)\|^2$ and the *Polyak–Łojasiewicz* inequality $2\mu\big(f(x)-f(x^\star)\big) \le \|\nabla f(x)\|^2$. Using the classical spectral characterization of GD on μ-SC, L_C-smooth quadratics, the worst-case linear rate is $\rho_C(\eta) = \max\{|1-\eta\mu|, |1-\eta L_C|\}$ for $\eta \in (0, 2/L_C]$, which bounds the distance contraction: $\|x^+ - x^\star\| \le \rho_C(\eta)\|x-x^\star\|$. Squaring and iterating K steps yields the claim. The bound $\rho_C(\eta) \le 1-\eta\mu$ holds since $|1-\eta L_C| \le 1-\eta\mu$ when $\eta \le 1/L_C$. □

Proof of Lemma G.30

Proof. Since $\nabla f_C(x_C^\star) = 0$, strong convexity of f_C gives

$$\mu\|x_C^\star - x^\star\| \;\le\; \|\nabla f_C(x^\star) - \nabla f_C(x_C^\star)\| \;=\; \|\nabla f_C(x^\star)\|.$$

Divide by μ to obtain the first inequality. Taking expectation over a uniformly sampled cohort C yields the second claim with $\sigma_\star^2 = \mathbb{E}\|\nabla f_C(x^\star)\|^2$. □

Proof of Proposition G.31

Proof. Fix the cohort C_t and denote $x_C^\star = \arg\min f_{C_t}$. By Lemma G.29,

$$\|x_{t,K} - x_C^\star\|^2 \;\le\; \rho_{C_t}^{2K}(\eta)\|x_t - x_C^\star\|^2.$$

Decompose $x_{t+1} - x^\star = (x_{t,K} - x_C^\star) + (x_C^\star - x^\star)$ and use $(a+b)^2 \le 2a^2 + 2b^2$:

$$\|x_{t+1} - x^\star\|^2 \;\le\; 2\|x_{t,K} - x_C^\star\|^2 + 2\|x_C^\star - x^\star\|^2 \;\le\; 2\rho_{C_t}^{2K}\|x_t - x_C^\star\|^2 + 2\|x_C^\star - x^\star\|^2.$$

Next expand $\|x_t - x_C^\star\|^2 \le 2\|x_t - x^\star\|^2 + 2\|x_C^\star - x^\star\|^2$ to obtain

$$\|x_{t+1} - x^\star\|^2 \;\le\; 2\rho_{C_t}^{2K}\big(2\|x_t - x^\star\|^2 + 2\|x_C^\star - x^\star\|^2\big) + 2\|x_C^\star - x^\star\|^2.$$

Collect terms of $\|x_C^\star - x^\star\|^2$:

$$\|x_{t+1} - x^\star\|^2 \;\le\; 4\rho_{C_t}^{2K}\|x_t - x^\star\|^2 \;+\; 2(1+2\rho_{C_t}^{2K})\|x_C^\star - x^\star\|^2.$$

Using Lemma G.30, $\|x_C^\star - x^\star\| \le \mu^{-1}\|\nabla f_C(x^\star)\|$, gives

$$\|x_{t+1} - x^\star\|^2 \;\le\; 4\rho_{C_t}^{2K}\|x_t - x^\star\|^2 \;+\; \frac{2(1+2\rho_{C_t}^{2K})}{\mu^2}\|\nabla f_{C_t}(x^\star)\|^2.$$

Renaming constants (absorbing factors of 2 into the front coefficients) yields the displayed bound in the statement. Finally, take expectation over the randomness of C_t and use $\mathbb{E}\|\nabla f_{C_t}(x^\star)\|^2 = \sigma_\star^2$ and $\rho_{C_t}(\eta) \le 1-\eta\mu$ to obtain the expectation bound. □